ASA-1169

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicants: S. NISHIOKA, et al

Serial No.: 10/790,062

Filing Date: March 2, 2004

For: INFORMATION SEARCHING METHOD, INFORMATION SEARCH SYSTEM, AND SEARCH SERVER

Art Unit: 2164

Examiner: A. M. Lewis

LETTER CLAIMING RIGHT OF PRIORITY

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

August 25, 2006

Sir:

Under the provisions of 35 USC 119 and 37 CFR 1.55, applicants hereby claim the right of priority based on:

Japanese Application No. 2003-104771
Filed: April 9, 2003

A Certified copy of said application document is attached hereto.

Acknowledgement thereof is respectfully requested.

Respectfully submitted,

Carl I. Brundidge
Registration No. 29,621
MATTINGLY, STANGER, MALUR & BRUNDIDGE, P.C.

CIB/jdc
Enclosures
703/684-1120

日本国特許庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されてる事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed th this Office.

出願年月日
Date of Application: ２００３年　４月　９日

出願番号
Application Number: 特願２００３－１０４７７１

[ST. 10/C]: [ＪＰ２００３－１０４７７１]

願人
Applicant(s): 株式会社日立製作所

CERTIFIED COPY OF
PRIORITY DOCUMENT

２００４年　３月　２日

特許庁長官
Commissioner,
Japan Patent Office

今井康夫

出証番号　出証特２００４－３０１５５７３

【書類名】　特許願

【整理番号】　H03004201A

【あて先】　特許庁長官　殿

【国際特許分類】　G06F 17/30

【発明者】

【住所又は居所】　東京都国分寺市東恋ケ窪一丁目２８０番地　株式会社日立製作所中央研究所内

【氏名】　西岡　真吾

【発明者】

【住所又は居所】　東京都国分寺市東恋ケ窪一丁目２８０番地　株式会社日立製作所中央研究所内

【氏名】　丹羽　芳樹

【発明者】

【住所又は居所】　東京都国分寺市東恋ケ窪一丁目２８０番地　株式会社日立製作所中央研究所内

【氏名】　今一　修

【特許出願人】

【識別番号】　000005108

【氏名又は名称】　株式会社　日立製作所

【代理人】

【識別番号】　100075096

【弁理士】

【氏名又は名称】　作田　康夫

【電話番号】　03-3212-1111

【手数料の表示】

【予納台帳番号】　013088

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　情報検索方法、情報検索システム及び検索サーバ

【特許請求の範囲】

【請求項１】

検索対象文字列が入力されるステップと、

前記検索対象文字列を用いて要約を作成し、要約単語リストを生成するステップと、

検索対象を絞り込むための制限条件が入力されるステップと、

前記要約単語リストに基づいて、文書データベースを用いて検索するステップと、

前記制限条件を付して、適合性を検査するステップと、

前記検索された結果でかつ前記制限条件に適合するものを、出力するステップとを有することを特徴とする情報検索方法。

【請求項２】

前記制限条件は、検索対象に必須な必須語または不要な不要語であることを特徴とする請求項１記載の情報検索方法。

【請求項３】

前記制限条件は、検索論理式であることを特徴とする請求項１記載の情報検索方法。

【請求項４】

制限条件に用いた語も連想検索のためのキーワードとして用いることを特徴とする請求項１記載の情報検索方法。

【請求項５】

前記適合性を検査するステップは、前記検索するステップで検索された検索結果について、前記制限条件を満たしているかどうかを判定することを特徴とする請求項１記載の情報検索方法。

【請求項６】

前記検索するステップは、前記制限条件を満たす文書を前記文書データベースから検索して適合性を検査した後、適合性を満たした検索結果と前記要約単語リ

ストとの類似度を比較して検索することを特徴とする請求項１記載の情報検索方法。

【請求項７】

前記要約単語リストを生成するステップは、前記検索対象文字列と選択された文書から生成することを特徴とする請求項１記載の情報検索方法。

【請求項８】

検索対象文字列を入力するための入力フレームと、

検索対象を絞り込むための制限条件を入力させるフレームと、

前記検索対象文字列を用いて要約を作成し、要約単語リストを生成する手段と、

前記要約単語リストに基づいて、情報検索装置の文書データベースを用いて検索させるための検索ボタンと、前記検索ボタンはその押下に応じて、前記情報検索装置に検索を指示するものであり、

前記文書データベースを用いて検索され、かつ前記制限条件を満たす検索結果を、前記情報検索装置から受け取り、その検索結果を表示する検索結果表示フレームとを有することを特徴とする情報検索システム。

【請求項９】

前記文書データベースを用いて検索させた検索結果を表示するフレームと、前記制限条件を入力させるフレームとは、同一の画面で並列に表示されることを特徴とする請求項８記載の情報検索システム。

【請求項１０】

情報検索を行う検索サーバであって、

文書に出現する単語と頻度から、その単語の重要度を計算する重要度計算手段と、

前記重要度計算手段にて計算された重要度の高い単語を候補として保持する、要約単語候補保持手段と、

検索要求の問い合わせ文章と選択された文書の要約として生成される要約単語リストは、前記重要度計算手段と前記要約単語候補保持手段によって、作成され、

前記要約単語リストと文書データベースに格納された文書との類似度を計算する類似度計算手段と、

前記文書データベースに格納された文書から、検索対象を絞り込むための制約条件に、適合するかどうかを検査する制約条件検査手段と、

前記類似度計算手段と前記制約条件検査手段とから検索された結果を保持する、検索結果候補保持手段とを有することを特徴とする検索サーバ。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は文書検索における情報検索方法、情報検索システム及び検索サーバに関する。

【０００２】

【従来の技術】

文書検索システムとして、従来から特開平１１－８５７８６号に記載されているように、与えられた文書や文章などに類似した文書をデータベース中から検索することが知られている。そのため、目的とする文書を的確にあらわすキーワードが思い付かない場合などでも、目的とする文書に近い文書をひとつでも発見できればその文書を指定し、連想検索を行うことで類似した文書を検索することが可能となる。

【０００３】

また、特開２００２－２２２２１０号には、類似文書型データベースとキーワード検索型データベースとを統合したメタサーチに関する連想検索システムが記載されている。このシステムでは、目的とする文書に近い文書が見付からない場合には、ある程度の長さの自然文を検索キーとして与えることも可能であり、自然文による検索システムであるともいえる。この機能により、執筆中の論文の一部やアブストラクト、また執筆中の特許などを与えることで、類似論文や類似特許を検索することも可能となっており、従来のキーワード検索とは大きく異なる。

【０００４】

連想検索システムは、与えられた文書や文書の一部、文章などに類似した文書をデータベース中から検索する。連想検索の実行には、与えられた文書等に現れる単語、文字などの頻度がしばしば用いられ、その結果、文書中の単語などを手がかりとして、類似した文書が検索される。この単語などを手がかりとした類似度の計算には統計的な手法が用いられ、その文書に現われる単語の分布などから内容的に類似した文書が検索される。

【特許文献１】　特開平１１－８５７８６号

【特許文献２】　特開２００２－２２２２１０号

【発明が解決しようとする課題】

連想検索システムにおいては通常、内容が統計的に類似度が高いものから順に結果が表示され、ユーザは類似度の尤度が高いものから選択的に検索結果を閲覧することができる。しかし、統計的に類似してたものが検索されるものの、文書の内容を理解しているわけではないので、ユーザの意図に忠実したがった検索は極めて困難である。

【０００５】

一方で、従来から広く用いられている、ＤＢ問い合わせ言語によるデータベースの検索や、その簡易なインタフェースで制約条件を指定してのデータベース検索、必須となるキーワードや禁忌キーワードを指定することによるデータベース検索では、連想検索とは異なり、その検索方法が表現可能な範囲内であればユーザの意図を厳密に反映することが可能である。しかし、これら従来の検索方法では、検索結果の表示順序はもしくはデータベースに依存した特定の順序で表示されるか、明示的に指定する場合でも何か特定のキーの値で整列させるしか無かった。

【０００６】

即ち、従来は、ユーザはいずれか一方の検索手段しか用いることができなかったのである。

【０００７】

【課題を解決するための手段】

そこで、本発明では、連想検索を行う際に、文書や文章に加えて対象となる文

書に対しての制約条件を与え、この制約を満たしつつ類似した文書の検索を行う構成とした。このようにすれば、ユーザの意図をより忠実に反映した連想検索を行うことが可能となり、より効率的な検索を行うことが可能となる。

【０００８】

以下、具体的に説明する。本願の連想検索システムは、連想検索を行うためのユーザインタフェース、連想計算サーバ、およびそれらの間の通信を媒介するネットワーク装置から構成される。

【０００９】

ユーザインタフェースは、連想検索の対象とする文書ＤＢを指定する手段と、連想検索の検索元となる文章を入力する手段と、連想検索の対象に適用すべき制約条件を入力する手段と、連想検索の開始を指示するボタンと、連想検索の結果を表示しかつ連想検索の検索元となる文書を指定する手段を有する。

【００１０】

連想検索サーバは、検索サーバプログラムを有しており、このプログラムは、図１１に示すように、重要度計算手段１０１１と、要約単語候補保持手段１０１２と、類似度計算手段１０１３と、制約条件検査手段１０１４と、検索結果候補保持手段１０１５を有し、検索対象となる文書ＤＢについて、各文書に現れる単語と、各単語が属する文書と、各文書についてのメタデータをあらかじめ解析して検索に用いることができるようにしている。

【００１１】

検索サーバは検索要求に類似した文書を検索するために、重要度計算手段および要約単語候補保持手段を用いて、検索元となる文書と文章の要約を作成し要約単語リストとする。つぎに、類似度計算手段により、要約単語リストに類似した文書を文書ＤＢから検索し、その類似度を、検索結果候補保持手段に保持されている最も類似度の小さな文書の類似度と比較することでその文書が検索結果の候補となりうるかを判定し、その場合、さらに制約条件検査手段により制約条件を満たしているかを判定し、そうであれば検索結果候補保持手段にその文書を加えることにより、制約条件を満たしつつ検索要求に類似した文書を検索する。このようにして、与えられた制約条件に適合するかどうかを検査し、その制約条件に

適合したものを、検索結果として出力する。

【００１２】

また、もうひとつの方法として、検索サーバは検索要求に類似した文書を検索するために、重要度計算手段および要約単語候補保持手段を用いて、検索元となる文書と文章の要約を作成し要約単語リストとする。つぎに、制約条件検査手段により制約を満たす文書を文書ＤＢから検索し、ついでその文書について要約単語リストとの類似度を計算し、その類似度を、検索結果候補保持手段に保持されている最も類似度の小さな文書の類似度と比較することでその文書が検索結果の候補となりうるか判定し、そうであれば検索結果候補保持手段にその文書を加えることにより、制約条件を満たしつつ検索要求に類似した文書を検索する。このようにして、与えられた制約条件に適合するかどうかを検査し、その制約条件に適合したものを、検索結果として出力する。

【００１３】

【発明の実施の形態】

（実施例１）

＜システム＞

以下、全体のシステムについて説明する。図１は本発明を実現するためのシステム構成例を示す概略図である。本システムは、ユーザからの検索要求を受付け、またはユーザに対して検索結果の表示を行うユーザインタフェース（２）、連想検索を実行する検索サーバ（１）、さらにそれらの間の通信を仲介する通信装置（９５）から構成される。通信装置（９５）は、計算機ハードウェア同士を接続しその間の通信を可能にするものである。

この通信装置は計算機ハードウェアおよびそれらを制御しているオペレーティングシステムから使用可能なものでなければならない。本発明ではこの通信装置として専用線による接続、ＬＡＮ、インターネットなどを使用する。

【００１４】

ユーザインタフェースおよび検索サーバは計算機ハードウェアおよびソフトウェアから構成されている。

【００１５】

なお、プラットフォームとして使用しているオペレーティングシステムが、単一のハードウェア上で複数のソフトウェアの実行を可能としている場合、本発明のユーザインタフェースと検索サーバは共にハードウェアを占有することなく動作可能であるため、単一のハードウェア上でユーザインタフェースおよび検索サーバを稼働させる構成が可能なのは明白である。この場合、ユーザインタフェースと検索サーバはオペレーティングシステムを介して通信することになり、ハードウェア間をつなぐ通信装置は不要となる（図３）。

【００１６】

なお、通信装置の使用の有無に関わらず、ユーザインタフェースと検索サーバはオペレーティングシステムを介して通信を行うわけであり、オペレーティングシステムが通信路の抽象化を行うため、いずれの場合も同一の動作を行うことで通信が可能である。従って、ソフトウェアの構成はいずれの場合でも同一にすることが可能である。そのため、以降、得に断らない限りユーザインタフェース、検索サーバなどのプログラム同士が通信をする場合について、通信装置を使う、使わないといった構成の違いには触れず、単に通信を行うと記す。

【００１７】

＜検索サーバ＞

検索サーバは、文書ＤＢの連想検索を実行するために文書ＤＢに関するデータ（４）を使用する。文書ＤＢに関するデータは、図７に示す通り、文書ＤＢに関するデータ（単語－文書）（４０１）、文書ＤＢに関するデータ（文書－単語）（４０２）、文書ＤＢに関するデータ（文書－メタデータ）（４０３）、からなる。文書ＤＢに関するデータ（文書－単語）（４０１）は、ＤＢ中の全ての文書について、その文書に現れる単語のＩＤと、その文書中でのその単語の出現頻度の組をリストにしたもので、あらかじめ作成しておく。文書ＤＢに関するデータ（単語－文書）は（４０２）文書ＤＢに関するデータ（文書－単語）（４０１）とは逆に、ＤＢ中の全ての単語についてその単語を含む文書のＩＤと、その文書中でのその単語の出現頻度の組をリストにしたもので、あらかじめ作成しておく。

なお、これは文書ＤＢに関するデータ（文書－単語）（４０１）の表全体を行列

ととらえ、その転置行列を作成することで容易に作成することができる。文書ＤＢに関するデータ（文書－メタデータ）（４０３）は、ユーザインタフェースで検索結果を表示するためのタイトルや検索結果の本文を表示するためのＵＲＬを含んでいる。これらのデータは他の通常の検索システムで使用するものと同じであり、文書データベースごとに適当なものを準備しなければならない。

【００１８】

なお、本文の表示が不要、またはユーザインタフェースのみで文書ＩＤなどを手がかりに本文の表示の実行が可能であったり、同じく文書ＩＤなどを手がかかりにタイトルの表示が可能である場合にはこの表の一部、または全部の作成が不要である。その場合、検索サーバがユーザインタフェースにこの表から取り出したデータを送る（後述）必要もない。

【００１９】

文書ＤＢに関するデータ（文書－メタデータ）（４０３）はあらかじめ作成しておく。

【００２０】

＜ユーザインターフェース＞

次に、ユーザインターフェースについて説明する。ユーザインタフェースのうち、ユーザから見える部分は、データベース選択部、問合せ入力部、制約条件入力部、検索開始ボタン、検索結果表示部から構成される。構成によっては文書連想検索開始ボタンを有することもある。

【００２１】

ユーザインタフェースの一例を図４に示す。データベース選択部（２１１）には、検索対象とするデータベースの名前を入力する。または選択肢が示され検索対象とするデータベースを選択することで検索対象となるデータベースをユーザインタフェースに伝えるものでもよい。これは、実装に用いたプラットフォームでの標準的な部品を用いることで実現可能である。

【００２２】

問合せ入力部（２１２）には連想検索の連想元となる文章、単語の列などを入力することができるようになっており、文章を入力することが可能な、ユーザイ

ンタフェースを実行するプラットフォームでの標準的な部品を用いることで実現可能である。

【００２３】

制約条件入力部（２１３）は、連想検索に際して追加して指定する制約条件を入力するための入力インタフェースである。これは、使用可能な制約条件に応じて異なる実装となる。

【００２４】

例えば、ＳＱＬの条件節を制約条件として使用可能な実装であれば、これは通常のテキスト入力インタフェース、もしくは、ＳＱＬの条件節の構造をパーズ、ハイライト表示できたりする構造エディタなどが使用可能である。必須単語や禁忌単語を制約条件として使用できるようにする場合には、それぞれに対応してテキスト入力窓を準備するか、単語の前に必須単語であれば“＋”、禁忌単語であれば“－”を付し、これら前置きされた記号を参照することで必須単語か禁忌単語を判別することも可能である。論理式などを用いる場合にも、テキスト入力インタフェースや、論理式用構造エディタなど、様々なインタフェースが使用可能である。本発明では、制約条件入力部としては、特別に実装したもの、もしくは標準的にそのシステムで使用可能であるもののいずれをも使用可能である。その意味で、制約条件入力部が本発明を適用しようとしている連想検索システムで実現・利用可能であること自明なことであり、また、前提であるともいえる。

【００２５】

検索開始ボタン（２１４）は、ユーザインタフェースを実行しているプラットフォームでの標準的な部品を用いることで実現されており、実装は容易である。

【００２６】

検索結果表示部（２１５）は主として検索結果をユーザに提示するために用いられ、検索結果である文書のタイトルや、検索要求に対する類似度などが表示される。また、本文を表示するための指示も検索結果表示部をとおして行うことができる。検索結果表示部のもうひとつの機能としては、表示されている複数の文書にマークをつけることである。このマークはユーザインタフェースから読み取ることが可能で、マークのつけられた文書は連想検索の際に、連想元文書として

扱われる。検索結果表示部はリストビューと呼ばれる標準部品などを用いることで容易に実装可能である。

【００２７】

ユーザインタフェースのうちユーザから直接操作できない部分はユーザインタフェースプログラム（２０１）、検索要求記憶部（２０２）、検索結果記憶部（２０３）からなる（図１）。

【００２８】

ユーザインタフェースプログラム（２０２）はユーザインタフェース全体を制御するための命令およびデータである。検索要求記憶部（２０２）は、ユーザが検索要求を問合せ入力部や制約条件入力部に入力したデータを一時的に記憶しておくためのものである。

【００２９】

検索結果記憶部（２０３）は、検索サーバから送り返されてきた検索結果をユーザインタフェースプログラムがユーザに提示するために、一時的に記憶しておくためのものである。

【００３０】

＜連想検索＞

図２９に、検索サーバで用いられる手続の依存関係の概要を示す。手続１が検索サーバによる連想検索のトップレベルであり、連想検索は手続１を呼び出すことによって行われる。手続２および手続５は、手続１から呼び出される。手続３は手続２から、手続４は手続３から、手続６は手続５から呼び出される。

【００３１】

図３０に、連想検索の実行にあたってこれらの手続の呼び出し・実行の状況を例示する。図で、縦軸は時間軸である。６１０の枠内はユーザインタフェースでの処理、６２０の枠内が検索サーバでの処理を示している。さらに、６２１の枠内は手続１の処理、６２２の枠内は手続２の処理、６２３の枠内は手続３の処理、６２４の枠内は手続４の処理、６２５の枠内は手続５の処理を示している。

【００３２】

ユーザインタフェースで検索開始のトリガがかかる（６１０１）と、ユーザイ

ンタフェースは検索要求を生成しそれを検索サーバに送る（６１０２）。検索サーバは手続１（６２１）を実行することでユーザインタフェースの要求を処理し、結果をユーザインタフェースに送り返す（６１０３）。なお、６１０３は検索結果表示を示す。手続１の実行（６２１）は、手続２の呼び出し（６２１１）、手続３の呼び出し（６２１２）、検索結果をユーザインタフェースに送り返す（６２１３は検索結果返信の準備と送信、６２１４は検索結果返信）という３つのステップからなる。

【００３３】

手続２（６２２）は、１：手続３を呼び出すことでユーザから送られてきた検索要求の要約を作成、２：検索単語リストを作成、３：要約単語リストと検索単語リストの併合、の３つのステップを実行する。要約単語リストの生成は主として手続３で行われる（６２３）。手続３は、１：文書リストの各文書ごとにその文書が含む単語リストを作成し、２：そのリストの要素のうち同一の単語を集めリストとしたものを引数として手続４を繰り返し呼び出し、３：要約単語候補保持手段（１０１２）の内容を出力、という３つのステップを実行する（６２３）。手続４は、手続３から渡された単語のリスト（全て同じ単語からなる）をもとに、その単語の重要度を重要度計算手段（１０１１）を用いて計算し、その結果を必要に応じて要約単語候補保持手段に蓄積する（６２４）。手続３が全ての単語について手続４を呼び出し終えた際に、要約単語候補保持手段に保持されているものが要約単語リストとなる。

【００３４】

手続５（６２５）は、手続２で作られた要約単語から、関連記事を検索する手続であり、１：要約単語リストの各単語ごとにその単語を含む文書リストを作成し、２：そのリストの要素のうち同一の文書を集めリストとしたものを引数として手続６を繰り返し呼び出し（６２５１）、３：検索結果候補保持手段（１０１５）の内容を出力、という３つのステップを実行する（６２５）。手続６（６２６）は、手続５から渡された文書のリスト（全て同じ文書からなる）をもとに、その文書の検索要求との類似度を類似度計算手段（１０１３）を用いて計算し、さらにその文書が制約条件を満たしているかを制約条件検査手段（１０１４）を

用いて検査し、その結果を必要に応じて検索結果候補手段に蓄積する（６２６）。手続５が全ての記事について手続６を呼び出し終えた際に、検索結果候補保持手段に保持されているものが検索結果となる。

【００３５】

この結果はユーザインタフェースに送り返され（６２１４）、検索結果としてユーザインタフェースによって表示される（６１０３）。以下では、それぞれについて詳細に説明する。

【００３６】

連想検索はマウス（９３２）などの物理的入力手段を用いて検索開始ボタン２１４（図４）を押し下げることで開始される。連想検索を開始する事象として、上記連想開始ボタン２１４を押し下げること以外に、キーボード（９３１、図１）などに備えられた改行キーや復帰キーが押し下げられることを用いても構わない。検索開始ボタン（２１４）、キーの押し下げなどがハードウェアおよびオペレーティングシステムによって感知されると、その事象はインタフェースプログラムに伝えられ、インタフェースプログラムは連想検索を開始する。連想検索が開始されると、最初に、インタフェースプログラムは、以下の手順で連想検索に必要な情報を収集する。

【００３７】

連想検索に必要な情報は、データベース選択部（２１１）（図４）によって選択された検索対象となるデータベース、問合せ入力部（２１２）に入力された問合せの文章や単語（以下、問合せ文章）、制約入力部（２１３）に入力された検索対象に対する制約、検索結果表示部（２１５）でマークされている文書（以下、連想元文書）である。インタフェースプログラムはこれらの情報を検索要求記憶部に記憶する。なお、文書につけたマークに基づいて文書からの連想を行う場合など、上述した情報全部を用いるのではなく、その一部だけ用いて連想検索を行うほうが好ましい場合もある。その場合、必要な情報のみを検索要求部に記憶しなければならない。すなわち、連想検索に必要な情報を選択的に収集する必要がある。これは、例えば図５のようにユーザインタフェースに文書連想検索開始ボタン（２１６）を追加することで検索開始のトリガとなるボタンを複数用意し

、押し下げられボタンに応じてあらかじめ準備しておいた図６のような対応表に基づき収集する情報を決定することで可能となる。

【００３８】

連想検索に必要な情報が収集されると、インタフェースプログラムはその情報を検索サーバに送る。検索サーバ（１）はこの情報（以下、検索要求）を受け取ると、指定された対象データベース中の文書で制約条件に合致するものに対して問合せ文章および連想元文書からの類似度を計算し（連想計算）その得点の高いものをインタフェースプログラムに送り返す。

【００３９】

これらの具体的手順は以下の通りである。

【００４０】

１．はじめに、図１２の手続１に示す通り、検索サーバは、検索要求にある対象データベースを特定し、ここ以降、その対象データベースについて検索を行えるよう、対象データベースへのアクセスを初期化する。

【００４１】

２．手続１に続き、図１２に示す通り、検索サーバは、手続２として、検索要求の中の問い合わせ文章と検索元文書から、要約単語リストを生成する。この手続２について具体的に説明する。まず、図１３に示す通り、検索要求の中の問合せ文章を単語に区切り、単語のリストを作る（５０１）。このリストを検索単語リストと呼ぶ。問合せ文章を単語に区切る作業は、形態素解析プログラム（１２）を使用することで容易に実行可能である。具体的には、形態素解析プログラムが起動されていなければ起動を行い、形態素解析プログラムとの通信路を確立する。通信路が確立された後、形態素解析すべき文字列をこの通信路を介して送り、ついで、それを解析して得られた形態素の列をこの通信路を介して受け取る。形態素解析が完了した後には、必要に応じて通信路を閉じ、最後に形態素解析プログラムを停止させる。以上の操作においてプログラムの起動、停止、通信などはオペレーティングシステムに要求することで容易に行える。形態素解析プログラムを外部プログラムとして呼び出すのではなく、検索サーバプログラムの一部として組み込むことも可能である。その場合には、オペレーティングシステムを

介しての通信は不要となり、プログラム内でデータを授受することができるのは明白である。

【００４２】

３．次に、手続３として、検索サーバは連想元文書を要約して、その文書を代表する単語のリストを作成する。この単語のリストには、単語ごとに重要度が実数で与えられている。検索サーバはその重要な順にあらかじめ定めてある適当な数mだけを取り出して使用する。この文書を代表する単語のリストを以後、要約単語リストと呼ぶ。

【００４３】

また、要約単語リストを作成するには、文書ＤＢに関するデータ（文書－単語）（４０１）を用いる。図８に示すように、文書ＤＢに関するデータ（文書－単語）（４０１）には、先に述べたように、各文書についてそれぞれの文書に出現する単語とその頻度が記録されている。

【００４４】

検索サーバは、まず、文書ＤＢに関するデータの表を照会することで、連想元文書すべてについてそれぞれに出現する単語とその頻度のリストを得る。すなわち、このリストは連想元文書の数と同じ個数得られる。これらのリストに現れる全ての単語についてその重要度を計算し、必要に応じて重要度の大きいものだけを取り出すことで要約とする。各単語の重要度の計算には統計的手法を用いることが適切であり、例えばよく知られているＴＦ・ＩＤＦ、ＳＭＡＲＴなどの尺度を用いることができる。もちろん、これもよく知られているものだが、超幾何分布に基づいた尺度や、ＳＭＡＲＴなどといったより高度な尺度を利用する場合でも、それを計算するモジュールを、定義式に対応させるだけで良いことはいうまでもない。

【００４５】

具体的に要約単語リストを作る手順は以下の通りである（手続３）。以下、図１４を参照して説明する。まず、連想元文書のリスト中の各文書について、文書ＤＢに関するデータ（文書－単語）を参照し、その文書を含む単語のリストを得る（５０３）。それぞれのリストの全ての要素に対してそのリストのキーとなる

文書をペアとし、すなわち単語とキーとなる文書と頻度の組を作成する。ついで、全ての組をひとつのリストにまとめ、全体を単語のＩＤで並べ替える。その結果に対してリストの先頭から順に見てゆき、同じ単語ＩＤが並んでいる間それらを集める。リストの次に来る要素が異なる単語ＩＤを持っていればそこまでで集めたものが同じ単語に関するものである。その単語は連想元文書リストのいずれかの文書に現れるもので、ここまでで集めたものがその連想元文書リストの文書のうちその単語を含む文書すべてである。この処理をリスト全体に渡って繰り返し行うことで、連想元文書リストに現れる文書に現れる単語のリストを作成することができる。なお、この処理はデータの整列・併合に関するもので古来より研究されている。そのため、これ以外にも様々な公知の方法があり、それらの方法を用いることも全く差し支えない。

【００４６】

また、次のステップ（５０４）では、この要約単語リストに対して順に処理を行ってゆく。即ち、単語のリストに現れる全ての単語について繰り返し処理を行う（５０４）。この作業では着目している要素のみが重要で、リスト全体は必要としない。そのため、このリストを作る作業と融合することで、リスト自体は作成せず、順に処理を行うことができるのは明白である。なんとならば、上記手順中でリストに加えるべき要素が得られた時点でこの作業を中断し、その要素に対して次のステップで行うべき作業（手続４）を行い、それが終了した後この作業を再開する、というように手順を変更するだけで実現可能だからである。

【００４７】

４．続いて手続４について、図１５を用いて説明する。検索システムは、上記の手順で作成した要約単語リストに現れる単語それぞれについて以下の手順を適用する。

【００４８】

まず、検索システムは着目している単語の重要度を計算する。そのために、まずこの単語を特徴付ける文書ベクトル、その単語を含む全ての文書のリストを取り出す。これは文書ＤＢに関するデータ（単語－文書）を参照することで容易に行うことができる。そして、その文書ベクトルと要約する文書リストとの類似度

すなわちこの単語の要約リストにおける重要度を計算することは、使用している類似度尺度に応じて定義されている数式にこれらふたつのベクトルをあてはめ、その値を計算するだけであり、それが先に述べたＴＦ・ＩＤＦなどの単純な尺度であれ、ＳＭＡＲＴや超幾何分布に基づく尺度など複雑なものであれ、計算そのものが容易であることは自明である。具体的には、この計算は、図１１に示すサーバの重要度計算手段（１０１１）によって行われる。

【００４９】

さて、さきにも述べたように、本発明は、要約のうちその大きい順にあらかじめ定めた個数mだけ、もしくは全部を使用する。全部を使用する場合には、とくに問題はなく、上記手順で作成された単語リストがそのまま要約となる。一方、m個だけ使用する場合には重要度の大きいm個だけをそのリストから取り出さなければならない。それには、完全なリストを作成し、リスト全体を重要度の大きい順に並べ替え、先頭からm個取り出す、という方法が最も自明な実装方法のひとつであり、この方法を用いても差し支えない。

【００５０】

別の方法として、以下のような方法がある（手続３および手続４の全体の動作）。この方法では、図１１に示すサーバの要約単語候補保持手段（１０１２）を用いる。

【００５１】

はじめに、要約単語候補保持手段を空にする。新しい単語の重要度が計算された時に、検索システムは条件に応じて以下の三通りの動作を選択する（５０５）。１つめとして、要約単語候補保持手段にm個未満の単語しかなければその単語を検索結果候補手段に追加する（５０６）。２つめとして、要約単語候補手段にm個の単語が含まれており、その中で最も小さい重要度より、現在処理中の単語の重要度が大きければ、要約単語候補保持手段から、最も小さな重要度を持つ単語を削除し、さらに現在処理中の単語を要約単語候補保持手段に追加する（５０７）。３つめとして、上記二通りのうちいずれにも該当しなければ検索システムはこの単語については何もしない。

【００５２】

さて、単語のリスト全てについて処理を完了した時点で、要約単語候補保持手段には、要約する文書リストでの重要度が大きいものから順にm単語、もしくは全体でm未満の単語しかなければ全ての単語が格納されている。これは、要約単語候補保持手段を順次更新した上記の手続より明らかであり、これが要約単語リストとなる。なお、文書の要約に際して、検索単語のリストももうひとつの文書として扱うことで、チェックされた文書と検索単語リストの要約単語リストを作ることも可能である。
この場合は、次に述べる検索単語のリストとの併合を行う必要はない。特に超幾何分布に基づいた尺度やＳＭＡＲＴなど高度な尺度を用いている場合にこの方法は特に有用である。

【００５３】

続いて、図１３の手順２に戻り、要約単語リストが計算できたら、次に検索サーバは、検索単語リストと要約単語リストとを併合する（５０２）。もちろん、検索の条件により、いずれか一方しか必要でない場合、もしくは検索単語のリストを要約文書として扱った場合であればこの作業は不要である。検索単語と要約単語リストとの併合の作業では、まず、それぞれの単語に付された重要度の擦り合わせを行う。要約単語リストの単語には重要度の計算に用いた尺度に基づいた値が付されているが、検索単語はその頻度しかえないため、たとえば、検索単語に付された頻度の最大の値を要約単語リストに現れる単語に付された最大の値と同じになるように値を調節する。また、いずれか一方を重視したい場合には、調節の後で、重視したい方の値を予め選んでおいた適当な定数倍するなどすることで容易に対応可能である。

【００５４】

重要度の擦り合わせが完了した後、ふたつのリストを連結し単語のＩＤ順に並べ替える。並べ替えはよく知られたソートアルゴリズムを使うことで容易に実行可能である。並べ替えたリストには検索単語と要約単語リスト中の単語とで重複するものが現れる可能性があるが、それらはこのリスト中では隣り合っているためリストを順に検査するだけで容易にそれらを検出可能である。重複する単語については重要度を加えあわせた上でひとつの単語としてリストを更新することで

、最終的には重複のないリストを作成することができる。

【００５５】

５．このようにして、手続２の要約単語リストの生成が完了したら、図１２に示すように、続いて手続５の要約単語リストから関連する文書を検索する工程に入る。検索サーバは併合された要約単語リストを文書とみなし、それに類似し、かつ検索要求の中の制約条件をみたす文書を検索対象ＤＢから検索する（手続５）。これらのうち類似度の高いものから順に要求された個数だけをユーザインタフェースに送り返す。これは、図１６に示すとおり以下の手順により行われる。

【００５６】

検索サーバは、要約単語リストに現れる単語を含む文書のリストを作成する。その際、各文書について要約単語リスト中の単語でその文書に含まれるもののリストも作成する。この手順では、文書ＤＢに関するデータ（単語－文書）（４０２）を用いる。

【００５７】

まず、要約単語リスト中の各単語について、文書ＤＢに関するデータ（単語－文書）を参照し、その単語を含む文書のリストを得る（５０８）。それぞれのリストの全ての要素に対してそのリストのキーとなる単語をペアとし、すなわち文書とキーとなる単語と頻度の組を作成する。ついで、全ての組をひとつのリストにまとめ、全体を文書のＩＤで並べ替える。その結果に対してリストの先頭から順に見てゆき、同じ文書ＩＤが並んでいる間それらを集める。リストの次に来る要素が異なる文書ＩＤを持っていればそこまでで集めたものが同じ文書に関するものであり、その文書に現れる単語でかつ要約単語リストにも現れる単語をペアに持つ組だけのリストとなり目的が達成される。この処理をリスト全体に渡って繰り返し行うことで、要約単語リストに現れる単語を含む文書のリストを作成することができる。なお、この処理はデータの整列・併合に関するもので、ひろく研究されており、これ以外にも様々な公知の方法があり、それらの方法を用いることも全く差し支えない。

【００５８】

また、次のステップ（５０９）ではこの文書のリストに対して順に処理を行っ

てゆくが、その作業では着目している要素のみが重要で、リスト全体は必要としない。そのため、このリストを作る作業と融合することで、リスト自体は作成せず、順に処理を行うことができるのは明白である。なんとならば、上記手順中でリストに加えるべき要素が得られた時点でこの作業を中断し、その要素に対して次のステップで行うべき作業を行い、それが終了した後この作業を再開する、というように手順を変更するだけで実現可能だからである。

【００５９】

次に、検索サーバは、要約単語リストの単語を含む文書のリストの各要素に対して、その文書を検索結果とするか否かを順次判定して行く（５０９）。まず、文書が検索問合せ中にある制約条件を満たすかどうかを検査する（５１０）。これには、検索サーバプログラムの一部である制約条件検査手段（１０１４、図１１）が用いられる。制約条件検査手段は、制約の種類に応じて以下に示した動作を行う。

【００６０】

まず、制約条件がＳＱＬの条件節である場合など、外部の手続を呼び出すことが適切である場合には外部のＤＢＭＳ（Ｄａｔａｂａｓｅ　Ｍａｎａｇｅｍｅｎｔ　Ｓｙｓｔｅｍ）などを呼び出すことで制約条件を満たすかどうかを確認する。具体的な手続は使用するＤＢＭＳおよびプラットフォームとして用いているオペレーティングシステムにより異なるが、それぞれの組み合わせごとに標準的方法が提供されており、そのシステムで適切な方法を使用することできわめて容易に実現可能であることは明白である。制約条件が、問合せ中の特定の単語についてそれが必須である、また、現れてはならない単語が指定されているなど、単語の存在に関するものである場合、処理中の文書についてその文書に現れる単語のリストは文書ＤＢに関するデータ（文書－単語）を参照することで容易に取り出すことができるため、条件に応じて特定の単語がその文書に現れているか否かを判定することはこれもまた容易である。複数の条件が連言として指定されている場合は全ての条件を満たさなければならないが、これは各条件を順に調べてゆきひとつでも不成立であれば条件を満たさないとし、全ての条件を調べ尽くした場合は成立とすればよい。複数の条件が選言として指定されている場合はいずれか

ひとつの条件を満たせば良いが、これは買う条件を順位調べてゆきひとつでも成立すればその時点で条件を満たすとし、全ての条件を調べ尽くした場合は不成立とすればよい。選言、連言の指定が入れ子になっている場合には外側の条件を処理している際には一段下の入れ子になっている部分をひとつの条件として処理を行えばよい。これは判定手続を再帰的に適用することに他ならず、入れ子が二段以上あってもこの方法で処理可能なことは数学的帰納法により簡単に確認できる。実装についても標準的な情報工学の手法を用いれば極めて容易である。

【００６１】

さて、制約条件検査手段によって条件判定が完了し、処理中の文書が問合せにある条件を満たしていない場合には、この文書は検索結果となり得ないから、単に無視して次の文書の処理に移れば良い。

【００６２】

処理中の文書が問合せにある条件を満たしている場合、この文書は検索結果になる可能性がある（５１１）。

【００６３】

６．このように、文書が制約条件を満たす場合は、図１７に示すように、手続６を行う。検索システムはこの要約単語リストとこの文書の類似度を計算する。そのために、まずこの文書を特徴付ける単語ベクトルを取り出す。これは文書ＤＢに関するデータ（文書－単語）を参照することで容易に行うことができる。そして、その単語ベクトルと要約単語リストとの類似度（すなわち要約単語リストと文書との類似度）を計算することは、使用している類似度尺度に応じて定義されている数式にこれらふたつのベクトルをあてはめ、その値を計算するだけであり、それが先に述べたＴＦ・ＩＤＦなどの単純な尺度であれ、ＳＭＡＲＴや超幾何分布に基づく尺度など複雑なものであれ、計算そのものが容易であることは自明である。この計算は類似度計算手段（１０１３、図１１）によって行われる。

【００６４】

さて、本発明の検索システムは、条件を満たす文書について類似度を付し、その大きい順に要求された個数（以下ｎ）もしくは全てを返すというものである。しかし、要約単語リストを作ったときと同様に、ある文書が類似度の大きい上位

ｎ文書に入っているかどうかを判定することは全ての文書の処理を終えるまで不可能であることに注意しなければならない。そこで、本発明では以下の方法によりこの検索結果リストを作成する。

【００６５】

すなわち、リストをある要素まで処理した時点での類似度が上位ｎまでに入っている文書を検索結果候補保持手段に貯えておくのである。ただし、それまでに処理した文書で条件を満たすものがｎ未満であった場合、検索結果候補保持手段にはそれら全て、すなわちｎ未満の文書を保持しておく。これは、以下の操作により具体的に実現することができる（手続５、手続６）。

【００６６】

新しい文書が結果の候補となった時に、検索システムは条件に応じて以下の三通りの動作を選択する。１つめに、検索結果候補保持手段にｎ個未満の文書しかなければその文書を検索結果候補手段に追加する（５１２）。２つめに、検索結果候補手段にｎ個の文書が含まれており、その中で最も小さい類似度より、現在処理中の文書の類似度が大きければ、検索結果候補保持手段から、最も小さな類似度を持つ文書を削除し、さらに現在処理中の文書を検索結果候補保持手段に追加する（５１３）。３つめに、上記二通りのうちいずれにも該当しなければ検索システムはこの文書については何もしない。

【００６７】

さて、文書のリスト全てについて処理を完了した時点で、検索結果候補保持手段には、条件を満たし、かつ要約単語リストとの類似度が大きいものから順にｎ文書、もしくは全体でｎ未満の文書しかなければ全ての文書が格納されている。これは、検索結果候補保持手段を順次更新した上記の手続より明らかであり、これが検索結果となる。むろん、要約単語リストを作成した際に説明したもうひとつの方法のように、全体のリストを作成し、それを類似度の大きい順に並べ替え、そこから上位ｎ文書取り出すという方法でも実現可能である。以上の手続により作成された検索要求に類似した文書のリストが検索結果である。

【００６８】

そして、検索システムは通信手段を介してこの検索結果をユーザインタフェー

スに送り返す。その際、必要に応じて文書のタイトル、ＵＲＬ等も検索結果に関連付けて送る。これはユーザインタフェースが結果の表示を行ったり、検索結果本文を取得する際に必要となるものである。なお、タイトル、ＵＲＬなどは文書ＤＢに関するデータ（文書－メタデータ）を参照することで容易に取得することができる。

【００６９】

なお、制限条件に用いた語も連想検索のためのキーワードとして用いることも、勿論可能である。

【００７０】

（実施例２）

前記実施例１では、手続５において、条件の判定を先に行い、それを満たすものについて類似度を計算し、それを検索結果候補保持手段に追加した。この順を逆にした実装について、本実施例２で説明する（図１８，１９）。

【００７１】

まず、図１８に示すように、手続７で、要約単語リストから各単語に対応して、その単語を含む文書のリストを作成し、続いて文書のリストに現れる全ての文書について繰り返す。この手順は、手続５と同様である。そして、図１９に示すように、文書のリストの要素を順に処理してゆく時点で、まず、処理中の文書と要約単語リストとの類似度を計算する。

【００７２】

そして、検索システムは条件に応じて以下の三通りの動作を選択する。１つめに、検索結果候補保持手段にｎ未満の文書しかない場合、処理中の文書が条件を満たすか否かを判定し、条件が満たされていれば検索結果候補保持手段にその文書を追加する（手続９、図２０）。２つめに、検索結果候補保持手段にｎ文書が保持されており、その中で最も小さな類似度より計算された類似度が大きい場合、処理中の文書が条件を満たすか否かを判定し、条件が満たされていれば検索結果候補保持手段中から最も小さな類似度を持つ文書を取り除き、次いで検索結果候補保持手段にその文書を追加する（手続１０、図２１）。３つめに、検索結果候補保持手段にｎ文書が保持されており、その中で最も小さな類似度より計算さ

れた類似度が小さい場合、その文書については何も行わない。

【００７３】

なお、このようにして作成した検索結果候補保持手段の最終的な内容が先に述べた方法で作成したものと同一になることは、処理の違いが、連言の処理の順番を入れ替えただけであり、ブール代数では交換則が成り立つことから明らかであろう。

【００７４】

（実施例３）

手続２ないし手続１０で使用した文書ＤＢに関するデータ（文書－単語）は単一の表であったが、この表は分割することも可能である。この表の分割した場合の処理を、連想元文書の要約を行う手続を例にして説明する。なお、要約単語リストから類似文書を検索する手続はここで説明する手続の文書と単語の役割を入れ替えたものとほぼ同一となり、その差分は制約条件の検査のみである。したがって、分割した表を用いて要約単語リストから類似文書を検索する手続はここで説明する手続を参考に極めて容易に実現可能であるから、その具体的な説明は割愛する。

【００７５】

図２２ないし図２５が、文書ＤＢに関するデータ（文書－単語）をふたつに分割した例である。図２２，図２４に示すとおり、その分割は、単語ＩＤの集合を適当な方法で互いに素なふたつの集合に分割し、それぞれの集合に含まれる単語ＩＤのみが各文書にあらわれるとみなし、それぞれの集合に対応したふたつの文書ＤＢに関するデータ（文書－単語）を作成している。この例では、第一の部分（図２３）には、単語ＩＤが１、２、...である単語のみが、第二の部分（図２５）には、単語ＩＤが３、４、...である単語のみが現れている。なお、各部分に対応する単語の集合は具体的に必要なものではなく、分割した表が矛盾なく作成できるのであれば、それが具体的に存在する、仮想的にしか存在しない、といったいずれの形態をとっても構わない。ここで仮想的というのは、表を分割する際に適当な手続が定義されており、その手続により分割された単語の集合が決定される、といった実現方法である。たとえば、２で割り切れるＩＤを持つ単語が

第一の集合に、２で割って１あまるＩＤを持つ単語が第二の集合に属する、という手続を定義することで、具体的に分割した単語集合を作成することなく、単語集合を定義することが可能である。

【００７６】

さて、一旦、分割した単語集合を作ることができれば、それぞれの集合に属する単語のみを要素として持つ、文書ＤＢに関するデータ（文書－単語）を分割したものを作成することが容易であることは自明であろう。なんとならば、一旦分割していない表を作成し、その表の各行について、単語集合に現れる要素のみを残した行を作成し、それを新しい表の対応する行にする、という手続を、それぞれの単語集合ごとに２回実行すれば、ふたつの分割された表を作成することができる。もちろん、全体の表を作成せず、いきなり分割した表を作ることも可能である。それには全体の表を作成する前に分割した単語集合を作成し、全体の表の各行を作成する際に分割した表それぞれに分配すればよい。この方法も実装は容易である。

【００７７】

さて、図２６に示す手続１１（５１４）（５１５）に示すように、このようにして分割した文書ＤＢに関するデータ（文書－単語）のそれぞれに対して、分割していないときと全く同様に、前述の要約単語リストを作成する手続を適用する（これが可能である理由は後述する）。この例の場合には、２分割であるから、２回適用することになる。なお、それぞれの適用はハードウェアやオペレーティングシステムが許すならば、並列に実行しても構わない。また、要約単語として使用する単語の数をｍとした場合、それぞれの適用ごとに、分割していないときと同様、ｍ個の要約単語を取り出す。

【００７８】

さて、分割している場合でも要約単語リストの作成が分割していない場合と同様に実行可能な理由は以下の通りである。表を分割した方法から直ちにに導けることだが、要約を行っている手続中である単語について処理を行っている際には、その単語を含む文書のリストは通常どおりに作成する方法と同一のものが得られている。この性質は要約単語から類似文書を検索する手続の場合に特に重要で

あり、すなわち、この性質により制約条件としてどのような必須単語の組が連言として与えられても正しく制約条件が検査できるのである。さらに、このようにして作成したそれぞれの要約単語リストは、その作り方から互いに疎であり、もし通常どおり作成した要約単語リストに対応単語があるとしたらその重要度は同じ値になっていることも保証される。

【００７９】

さて、手続１２（図２７）に示すとおり（これは、図２６手続１１（５１６）から呼び出される）、それぞれの部分について作成されたふたつの要約単語リストを併合し、重要度が大きい順にm個をとれば、通常どおり作成した要約単語リストに一致することは、上述した理由からも明らかである。このようにして、分割した表を用いて要約単語リストを作成することが可能である。なお、それぞれの部分に対する要約の作業でm個の結果が必要だと述べたが、これをｍより小さい値、例えばｋにすることも可能である。この場合、併合を行った結果、m個の単語を得るためにｋはｍ／２以上でなければならないことは当然であるが、そうであったとしても、併合の際にどちらか一方での結果のｋ個を全て使いきってしまった場合、その最小の重要度をもつ単語より重要度が小さい単語については通常の方法で作成した要約単語リストに現れるのにこの方法で作成したリストには現れない可能性がある（結果から漏れてしまう、別の言い方をすれば他のものが入ってくる）。

【００８０】

単語の集合の分割を十分ランダムに行えば、この状態が発生する確率は二項分布の累積確率の計算式を用いて計算可能であり（図２８）、適当なｋを選ぶことによって、確率的に任意の精度を実現することが可能である。

【００８１】

この例では表を２分割したが、一般に何分割しても構わない。その方法は２分割した方法に準じて分割数を増やすだけであり、実現は２分割の場合と同程度に容易である。分割した各部分に対応して要約単語リストを作成する作業は検索サーバと異なるハードウェア上で動作しているプロセス、もしくは同一ハードウェア上で動作している異なるプロセスで行うことも可能である。その際、それぞれ

のプロセスは少なくともそれぞれのプロセスで要約を行う分割された表へのアクセスが可能でなければならない。さらに文書連想の場合で制約条件が外部プログラムで行われる場合には通常の場合と同様のその外部プログラムへのアクセスが可能でなければならない。いずれも、運用の際に表を記憶させるハードウェアの設定、オペレーティングシステム、その他ＤＢＭＳの設定を適切におこなえば容易に実現できることである。さらに、検索サーバがそれぞれの分割された部分の要約を行うプロセスと通信することができなければならない。これは、オペレーティングシステムを介して通信させるものとし、通信相手として特定のプロセスを指定する方法も、使用しているプラットフォームで提供される方法が利用可能であるから、その実現に関してなんらの困難も存在しない。

【００８２】

図２に２分割し、ハードウェアも２セット使用したシステムの構成例を示す。この例では検索サーバ（正）（１０４１）がユーザインタフェースとのやり取り、第一の部分に対する要約・文書連想、各部分の要約結果の併合を担当しており、検索サーバ（副）（１０４２）が第二の部分に対する要約・文書連想を担当している。特に図示はしないが、検索サーバ（副）を増やし、それぞれが要約・文書連想を担当し、検索サーバ（正）がユーザインタフェースとのやり取り、各部分の要約結果の併合を担当するという構成が可能なのはいうまでもない。また、分割数を２より大きくした場合でも、検索サーバ（副）を分割数に応じて増加させるだけで良い。

【００８３】

なお、いずれの例でも、文書連想を行うにあたり単語を用いたものを示したが、単語ではなく、文字ｎ－ｇｒａｍ、塩基配列ｎ－ｇｒａｍ、アミノ酸二次構造を適当な長さに区切ったものなど、文書を特徴付けるもの（特徴素）であれば使用可能である。この場合、形態素解析プログラムをこのような特徴素に対応したものに置き換えることで、容易にこれらの特徴素に対応可能であるということはいうまでもない。

【００８４】

【発明の効果】

ユーザは、連想検索を行う際に、文書や文章に加えて対象となる文書に対しての制約条件を与えることができる。検索システムは、この制約を満たし、かつ、キーとなる文書や文章に類似した文書の検索を行う。結果は従来の連想検索と同様、尤度の高いものから順に表示される。その結果、ユーザの意図をより忠実に反映した連想検索を行うことが可能となり、より効率的な検索を行うことが可能となる。

【図面の簡単な説明】

【図１】

本発明の実施例の構成を示す図。

【図２】

文書ＤＢに関するデータを２分割した場合の構成図。

【図３】

ユーザインタフェースおよび検索サーバを同一ハードウェアで動作させた場合の構成図。

【図４】

ユーザインタフェースの例。

【図５】

連想検索開始ボタンを有するユーザインタフェースの例。

【図６】

ボタンごとに収集すべきデータを記載した表。

【図７】

文書ＤＢに関するデータ。

【図８】

文書ＤＢに関するデータ（文書－単語）。

【図９】

文書ＤＢに関するデータ（単語－文書）。

【図１０】

文書ＤＢに関するデータ（メタデータ）。

【図１１】

検索サーバプログラム。

【図１２】

手続１：検索サーバの動作。

【図１３】

手続２：要約単語リストの作成。

【図１４】

手続３：要約単語リストの作成（本体）。

【図１５】

手続４：要約単語リストの作成（重要度の計算と、要約単語候補保持手段の更新）。

【図１６】

手続５：要約単語リストから関連する文書を検索。

【図１７】

手続６：要約単語リストから関連する文書を検索（本体）。

【図１８】

手続７：要約単語リストから関連する文書を検索（別法）。

【図１９】

手続８：要約単語リストから関連する文書を検索（別法、本体）。

【図２０】

手続９：制約条件の検査と、検索結果候補保持手段への単語の追加。

【図２１】

手続１０：制約条件の検査と、検索結果候補保持手段の単語の更新、

【図２２】

文書ＤＢに関するデータの一部、第一の部分。

【図２３】

文書ＤＢに関するデータ（文書－単語）第一の部分。

【図２４】

文書ＤＢに関するデータの一部、第二の部分。

【図２５】

文書ＤＢに関するデータ（文書－単語）第二の部分。

【図２６】

手続１１：文書を要約して要約単語リストを作る（分割版）。

【図２７】

手続１２：ふたつの要約単語リストの併合。

【図２８】

完全な要約・連想結果が得られなくなる確率の上限ｅの計算式。

【図２９】

手続の依存関係。

【図３０】

手続実行の例。

【符号の説明】

１:検索サーバ

１０１:検索サーバプログラム

１０１１:重要度計算手段

１０１２:要約単語候補保持手段

１０１３:類似度計算手段

１０１４:制約条件検査手段

１０１５:検索結果候補保持手段

１０４１:連想検索サーバ（正）

１０４２:連想検索サーバ（副）

１１:ＤＢＭＳ検索エンジン

１２:形態素解析プログラム

２:ユーザインタフェース

２０１:ユーザインタフェースプログラム

２０２:検索要求記憶部

２０３:検索結果記憶部

２０４:ボタンごとの送信するデータの表

２１１:データベース選択部

２１２:問合せ入力部

２１３:制約条件入力部

２１４:検索開始ボタン

２１５:検索結果表示部

２１６:文書連想検索開始ボタン

４:文書ＤＢに関するデータ

４０１:文書ＤＢに関するデータ（文書－単語）

４０２:文書ＤＢに関するデータ（単語－文書）

４０３:文書ＤＢに関するデータ（文書－メタデータ）

４１０:文書ＤＢに関するデータの一部（第一の部分）

４１１:文書ＤＢに関するデータ（文書－単語）第一の部分

４２０:文書ＤＢに関するデータの一部（第二の部分）

４２１:文書ＤＢに関するデータ（文書－単語）第二の部分

５０１:検索単語リスト作成

５０２:検索単語リストと要約単語リストの併合

５０３:文書に現れる単語リストを作成

５０４:要約単語リスト作成の本体

５０５:要約単語候補保持手段を更新するか否かの判断

５０６:要約単語候補保持手段への単語の追加

５０７:要約単語候補保持手段にある単語の差し替え

５０８:単語リストに現れる文書リストの作成

５０９:連想検索の本体

５１０:条件判断の実行

５１１:条件判断が真となった場合

５１２:検索結果候補保持手段への単語の追加

５１３:検索結果候補保持手段にある単語の差し替え

５１４：問合せ文書のリストを要約して要約単語リストを作成（第一の部分）

５１５：問合せ文書のリストを要約して要約単語リストを作成（第二の部分）

５１６：要約の併合

６１０：ユーザインタフェースの処理

６１０１：検索開始トリガ

６１０２：検索要求送信

６１０３：検索結果表示

６２０：検索サーバの処理

６２１：手続１の処理

６２１１：手続２の呼び出し

６２１２：手続３の呼び出し

６２１３：検索結果返信の準備と送信

６２１４：検索結果返信

６２２：手続２の処理

６２３：手続３の処理

６２４：手続４の処理

６２５：手続５の処理

６２５１：手続６の呼び出し

６２６：手続６の処理

９１:制御・演算装置

９２:記憶装置

９３:入力装置

９３１:キーボード

９３２:マウス

９４:出力装置

９４１:ディスプレイ

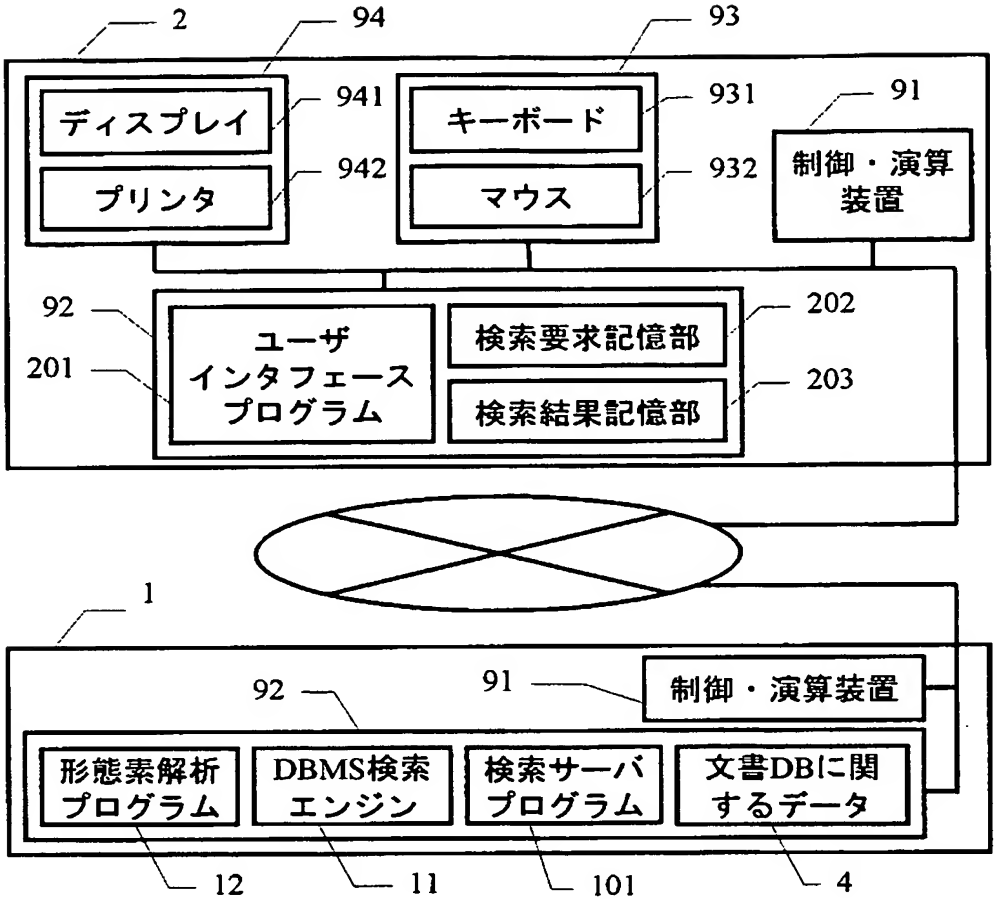
９４２:プリンタ

９５:通信装置。

【書類名】　図面

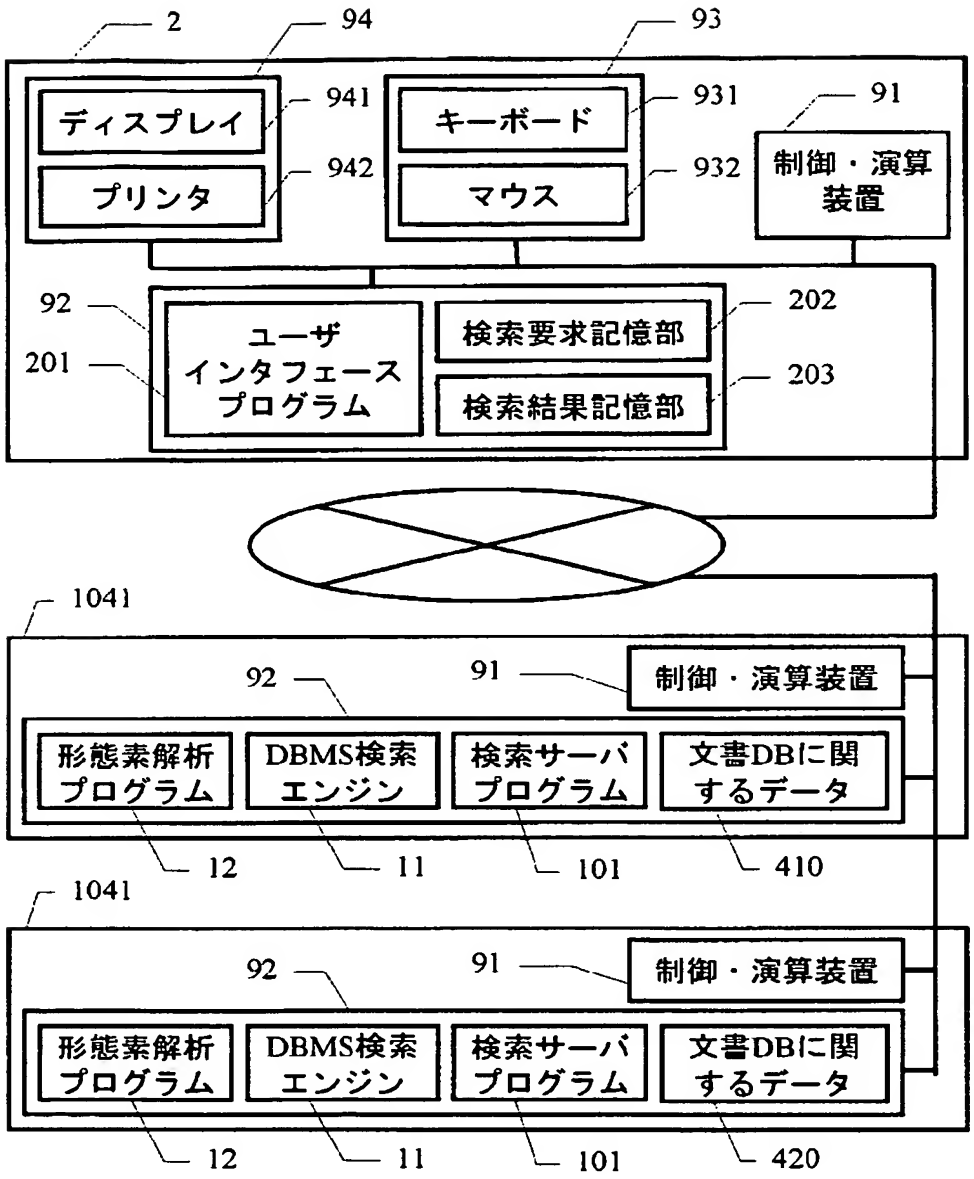
【図１】

図1


2
94
93
ディスプレイ
941
キーボード
931
91
プリンタ
942
マウス
932
制御・演算装置
92
ユーザインタフェースプログラム
201
検索要求記憶部
202
検索結果記憶部
203
1
92
91
制御・演算装置
形態素解析プログラム
DBMS検索エンジン
検索サーバプログラム
文書DBに関するデータ
12
11
101
4

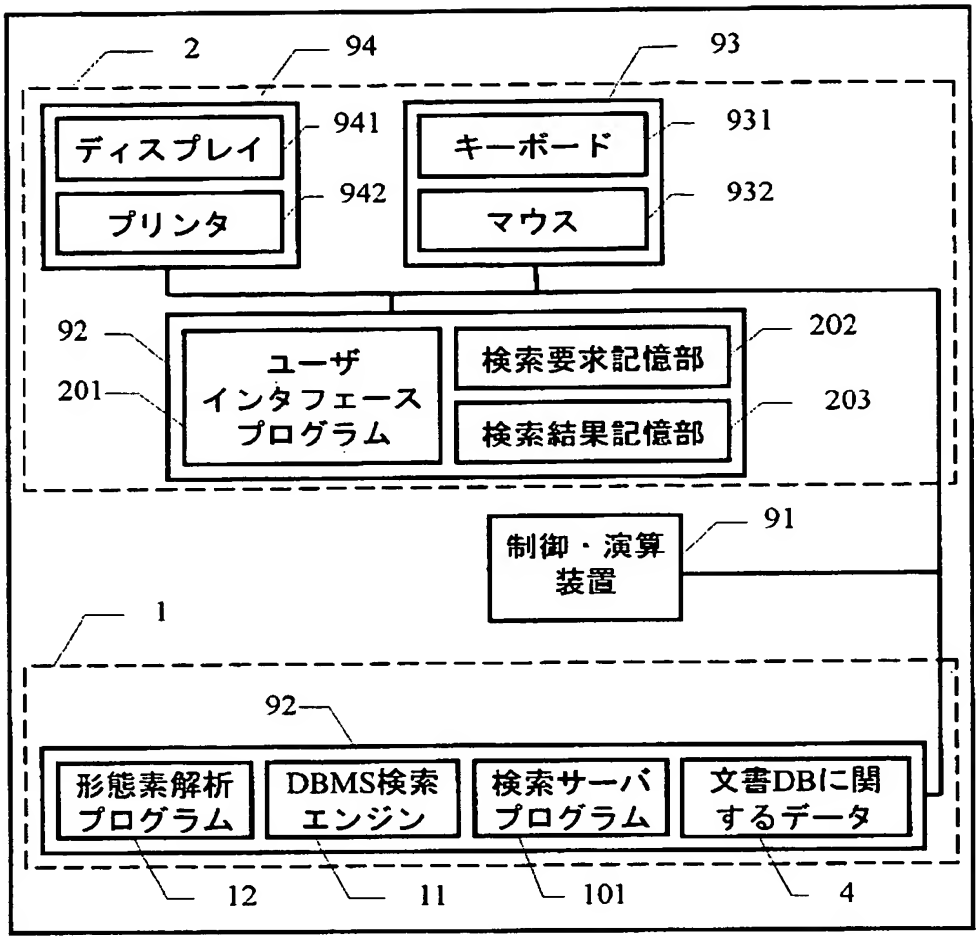
【図２】

図2


2
94
93
ディスプレイ
941
キーボード
931
91
プリンタ
942
マウス
932
制御・演算装置
92
ユーザインタフェースプログラム
201
検索要求記憶部
202
検索結果記憶部
203
1041
92
91
制御・演算装置
形態素解析プログラム
DBMS検索エンジン
検索サーバプログラム
文書DBに関するデータ
12
11
101
410
1041
92
91
制御・演算装置
形態素解析プログラム
DBMS検索エンジン
検索サーバプログラム
文書DBに関するデータ
12
11
101
420

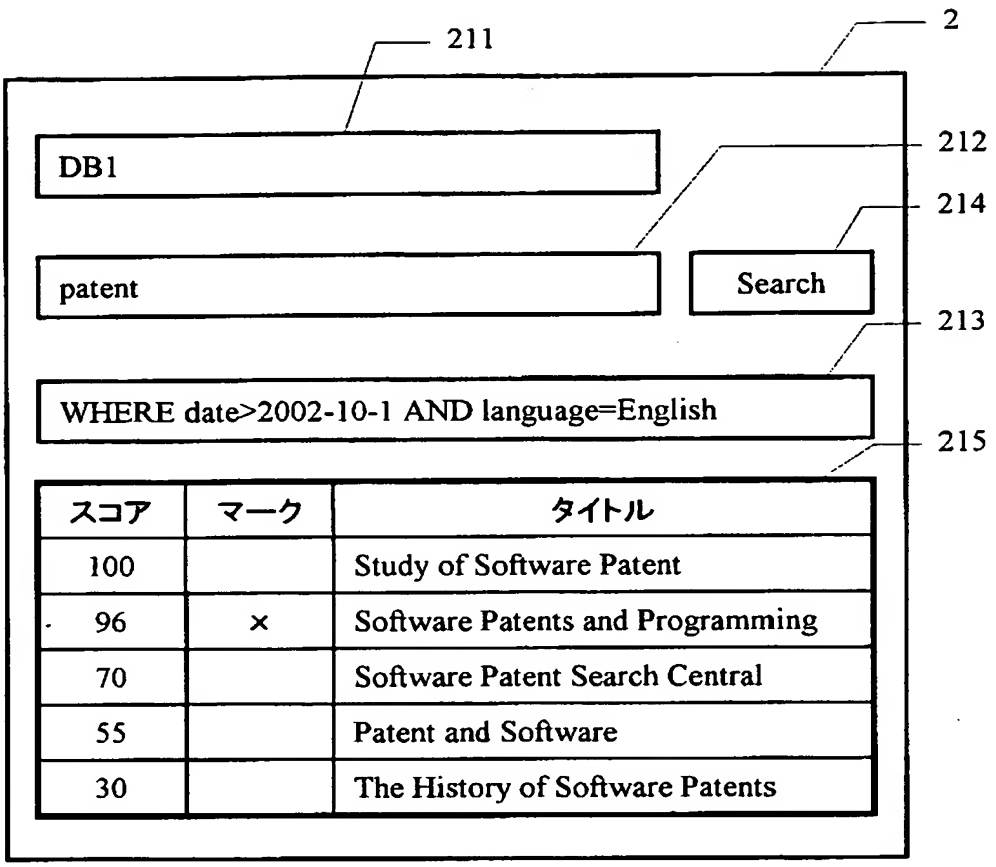
【図３】

図3


2
94
93
ディスプレイ
941
キーボード
931
プリンタ
942
マウス
932
92
201
ユーザ
インタフェース
プログラム
検索要求記憶部
202
検索結果記憶部
203
制御・演算
装置
91
1
92
形態素解析
プログラム
DBMS検索
エンジン
検索サーバ
プログラム
文書DBに関
するデータ
12
11
101
4

【図４】

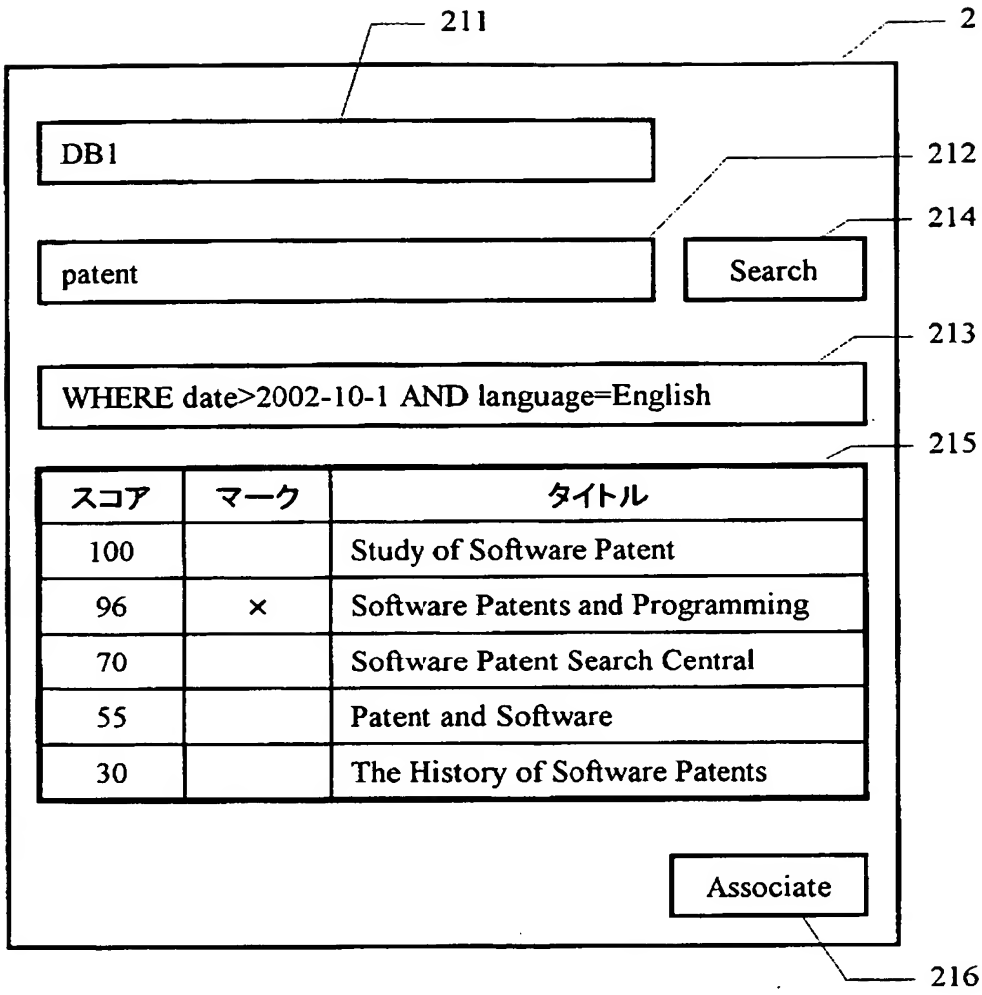
図4



スコア	マーク	タイトル
100		Study of Software Patent
96	×	Software Patents and Programming
70		Software Patent Search Central
55		Patent and Software
30		The History of Software Patents

【図５】

図5


211
2
DB1
212
214
patent
Search
213
WHERE date>2002-10-1 AND language=English
215
スコア
マーク
タイトル
100
Study of Software Patent
96
×
Software Patents and Programming
70
Software Patent Search Central
55
Patent and Software
30
The History of Software Patents
Associate
216

【図６】

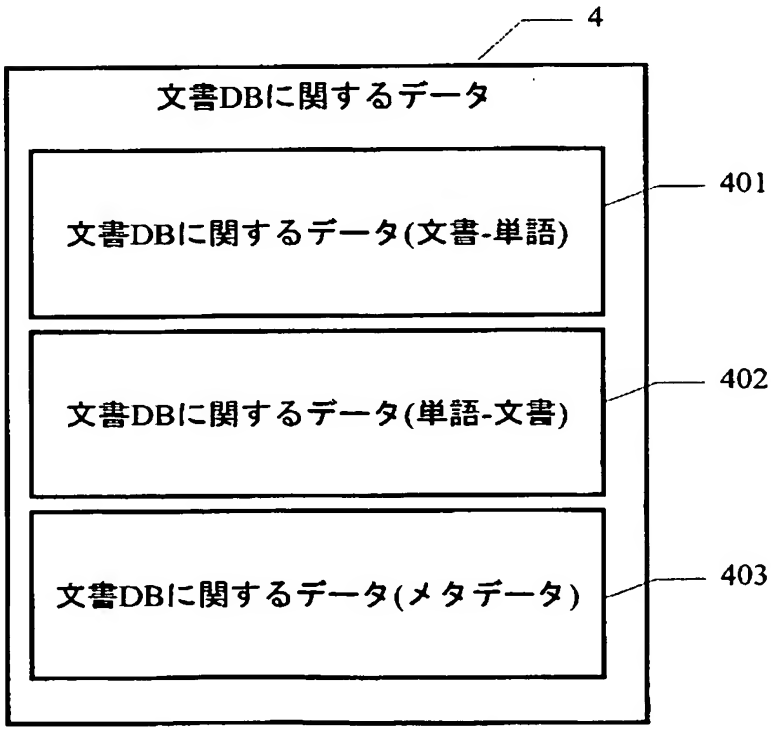
図6

204

押し下げられたボタン	収集する情報
検索ボタン	対象DB 検索要求入力部 検索条件入力部 文書のチェック
文書連想検索ボタン	対象DB 文書のチェック

【図７】

図7



【図８】

図8

401

文書ID	含まれる単語IDと頻度の対
1	{1, 10}, {2, 3}, {4, 8}, ...
2	{2, 5}, ...
3	{4,2},...
4	...
...	...

【図９】

図9

402

単語ID	含む文書IDと頻度の対
1	{1, 10}, ...
2	{1,3}, {2, 5}, ...
3	...
4	{1,8}, {3,2}, ...
...	...

【図１０】

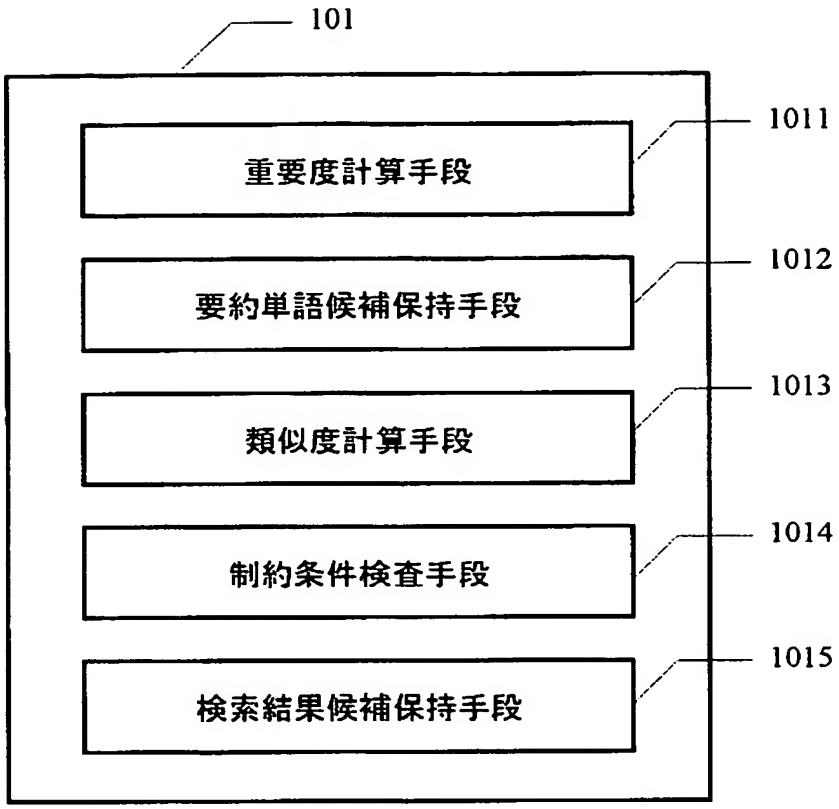
図10

403

文書ID	タイトル	URL
1	Patent and Software	http://www.some.where/
2	The History of Software Patents	ftp://ftp.now.here/
3	Software Patent Institute	...
4	...	
...	...	

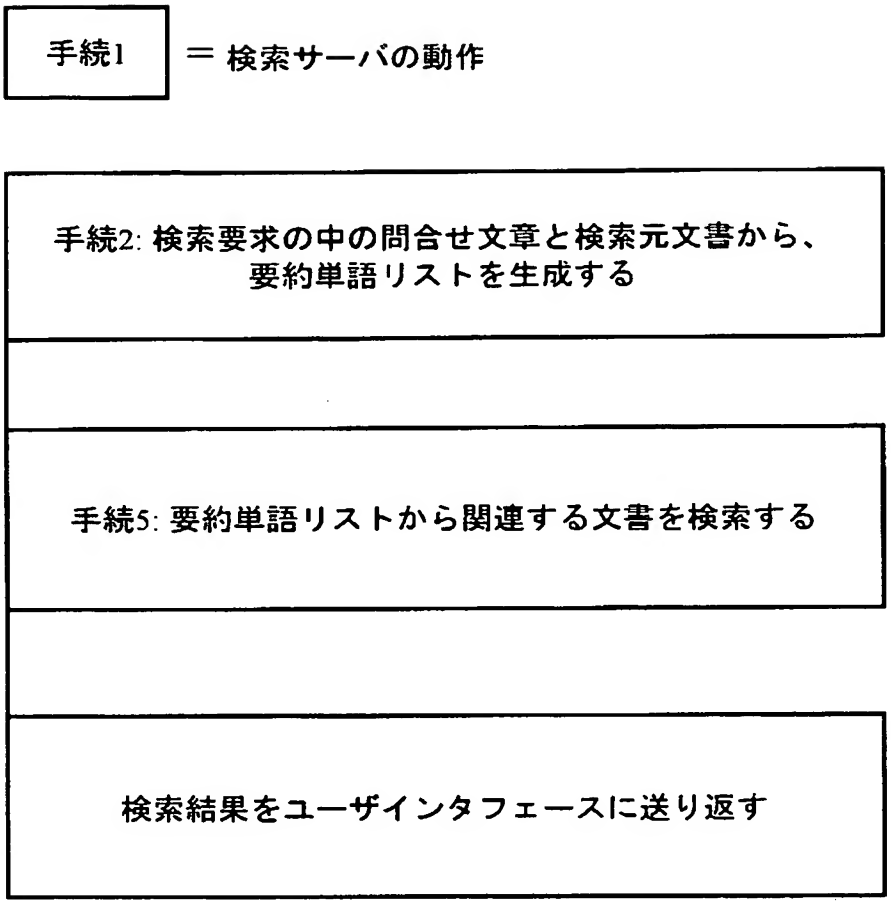
【図１１】

図11


101
1011
重要度計算手段
1012
要約単語候補保持手段
1013
類似度計算手段
1014
制約条件検査手段
1015
検索結果候補保持手段

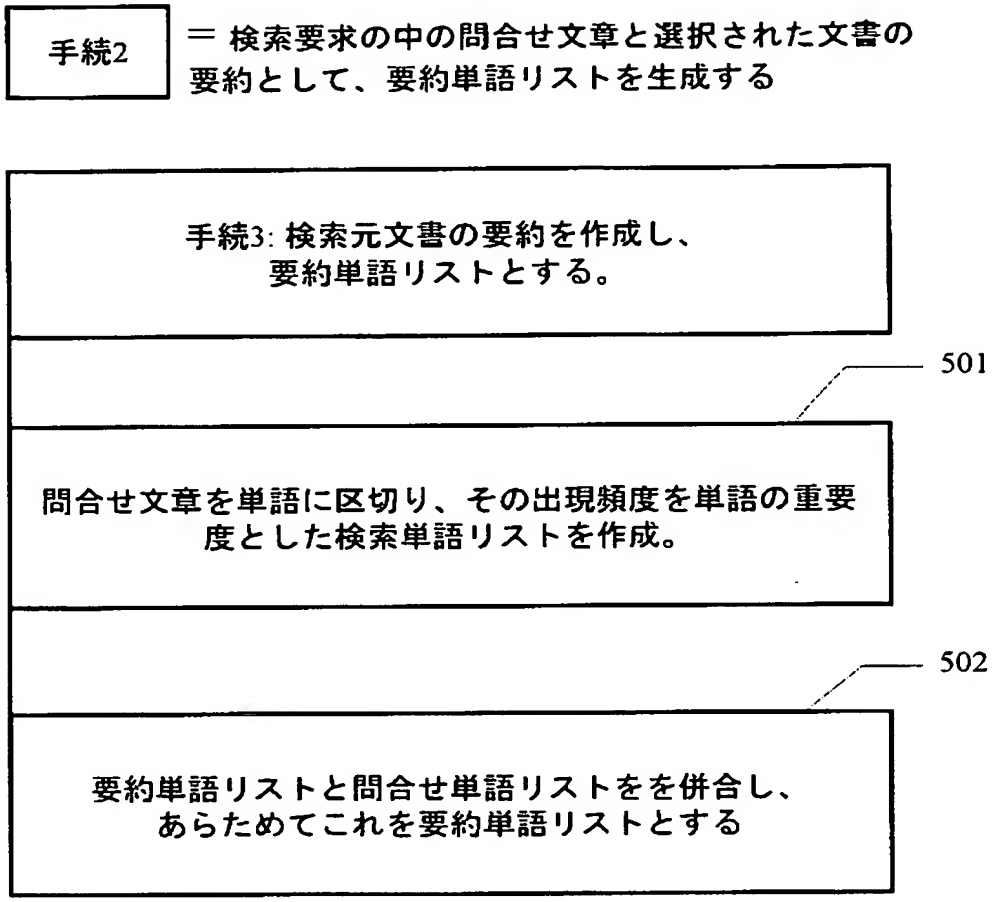
【図１２】

図12


手続1
＝ 検索サーバの動作
手続2: 検索要求の中の問合せ文章と検索元文書から、
要約単語リストを生成する
手続5: 要約単語リストから関連する文書を検索する
検索結果をユーザインタフェースに送り返す

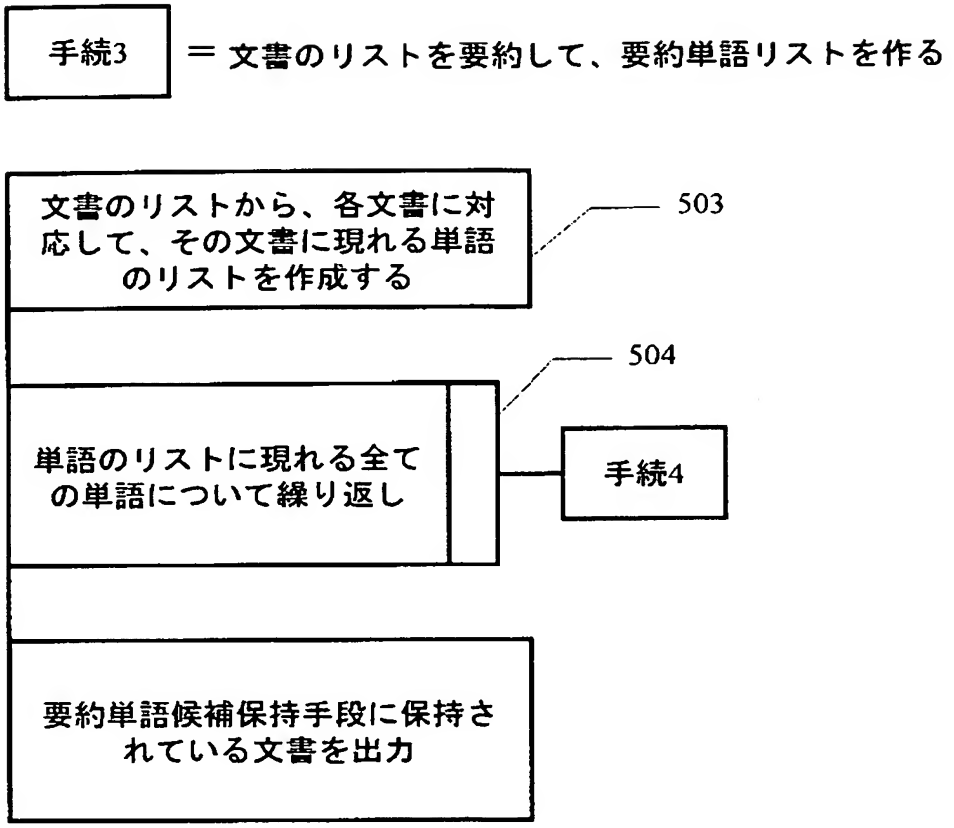
【図１３】

図13


手続2
＝検索要求の中の問合せ文章と選択された文書の要約として、要約単語リストを生成する
手続3: 検索元文書の要約を作成し、要約単語リストとする。
501
問合せ文章を単語に区切り、その出現頻度を単語の重要度とした検索単語リストを作成。
502
要約単語リストと問合せ単語リストをを併合し、あらためてこれを要約単語リストとする

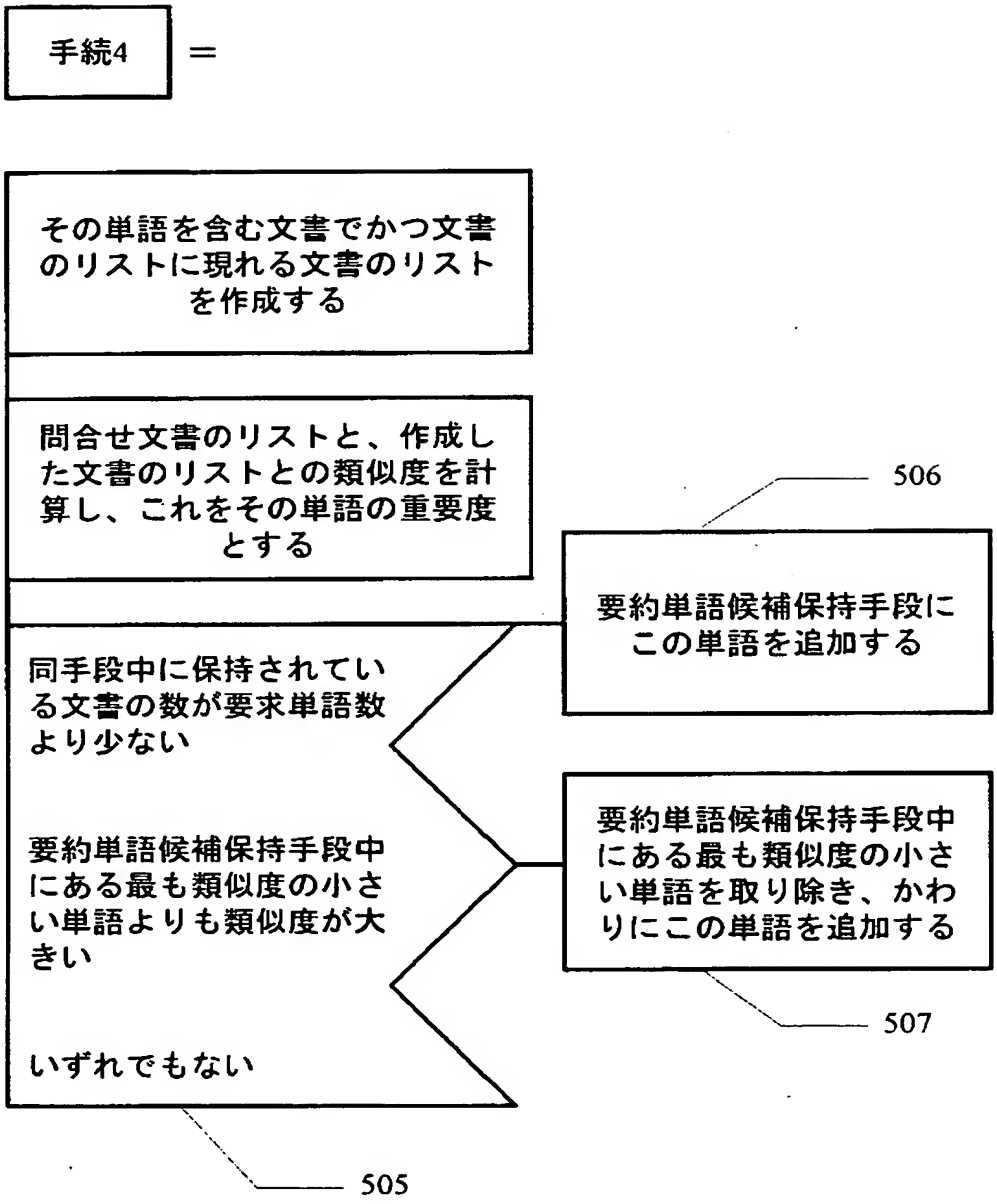
【図１４】

図14


手続3
＝ 文書のリストを要約して、要約単語リストを作る
文書のリストから、各文書に対応して、その文書に現れる単語のリストを作成する
503
504
単語のリストに現れる全ての単語について繰り返し
手続4
要約単語候補保持手段に保持されている文書を出力

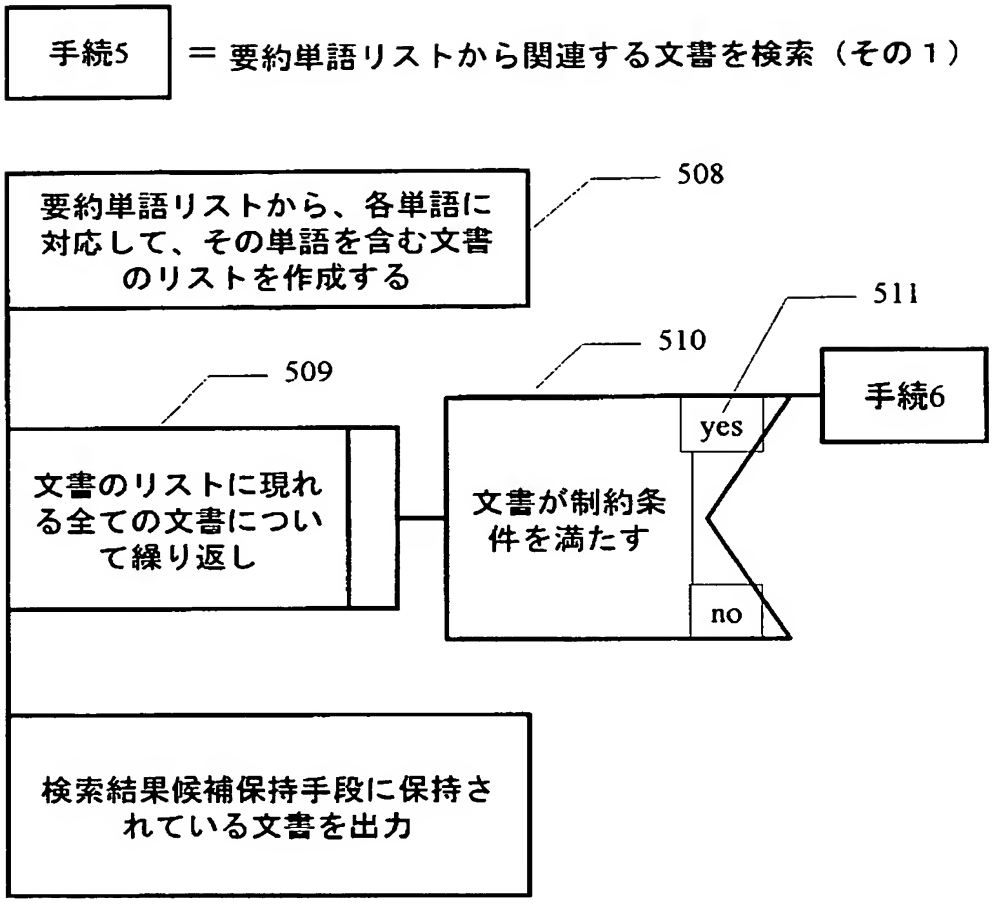
【図１５】

図15


手続4
＝
その単語を含む文書でかつ文書のリストに現れる文書のリストを作成する
問合せ文書のリストと、作成した文書のリストとの類似度を計算し、これをその単語の重要度とする
同手段中に保持されている文書の数が要求単語数より少ない
要約単語候補保持手段中にある最も類似度の小さい単語よりも類似度が大きい
いずれでもない
505
506
要約単語候補保持手段にこの単語を追加する
507
要約単語候補保持手段中にある最も類似度の小さい単語を取り除き、かわりにこの単語を追加する

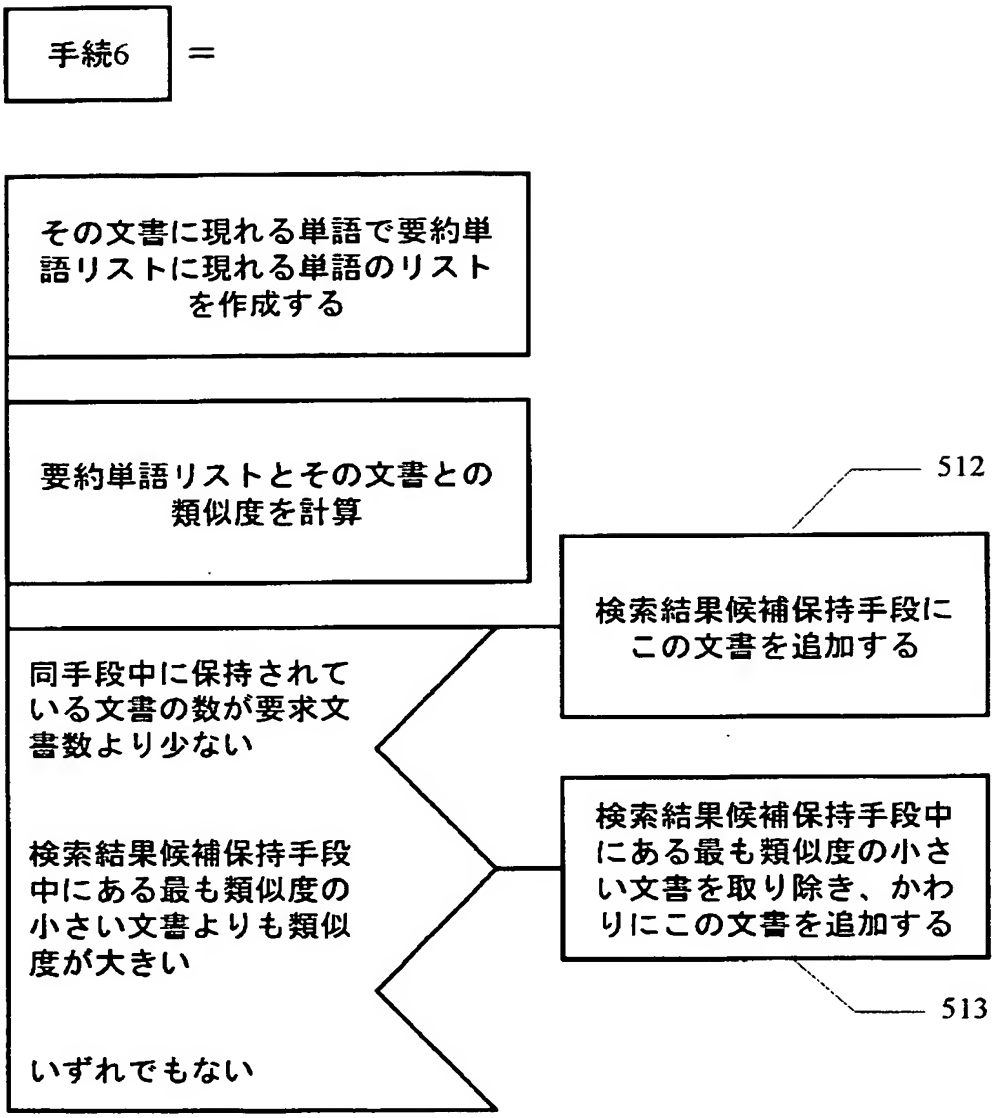
【図１６】

図16


手続5
＝ 要約単語リストから関連する文書を検索（その１）
要約単語リストから、各単語に対応して、その単語を含む文書のリストを作成する
508
511
510
509
手続6
yes
文書のリストに現れる全ての文書について繰り返し
文書が制約条件を満たす
no
検索結果候補保持手段に保持されている文書を出力

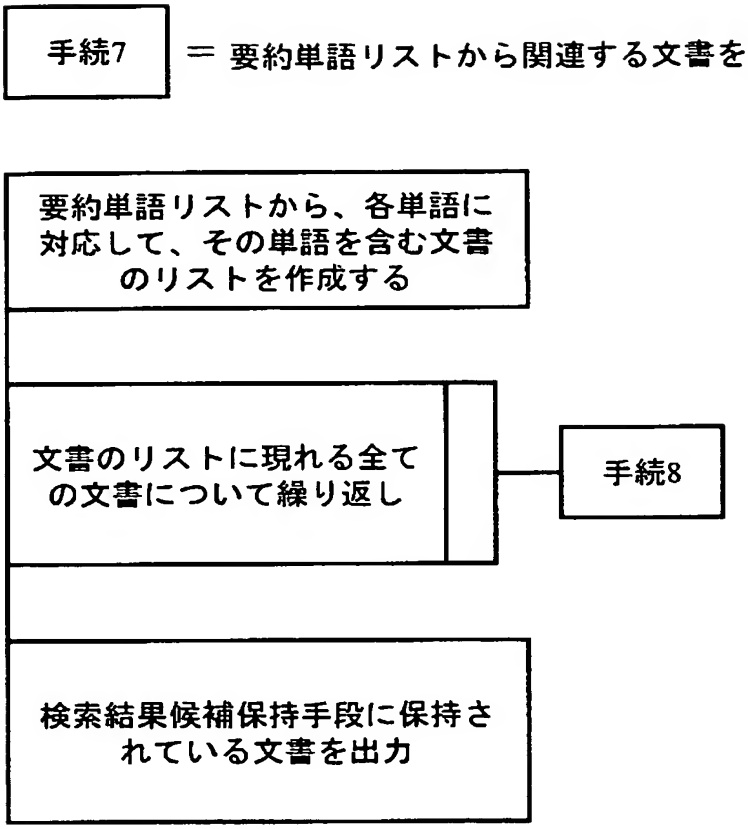
【図１７】

図17


手続6
＝
その文書に現れる単語で要約単語リストに現れる単語のリストを作成する
要約単語リストとその文書との類似度を計算
512
検索結果候補保持手段にこの文書を追加する
同手段中に保持されている文書の数が要求文書数より少ない
検索結果候補保持手段中にある最も類似度の小さい文書を取り除き、かわりにこの文書を追加する
検索結果候補保持手段中にある最も類似度の小さい文書よりも類似度が大きい
513
いずれでもない

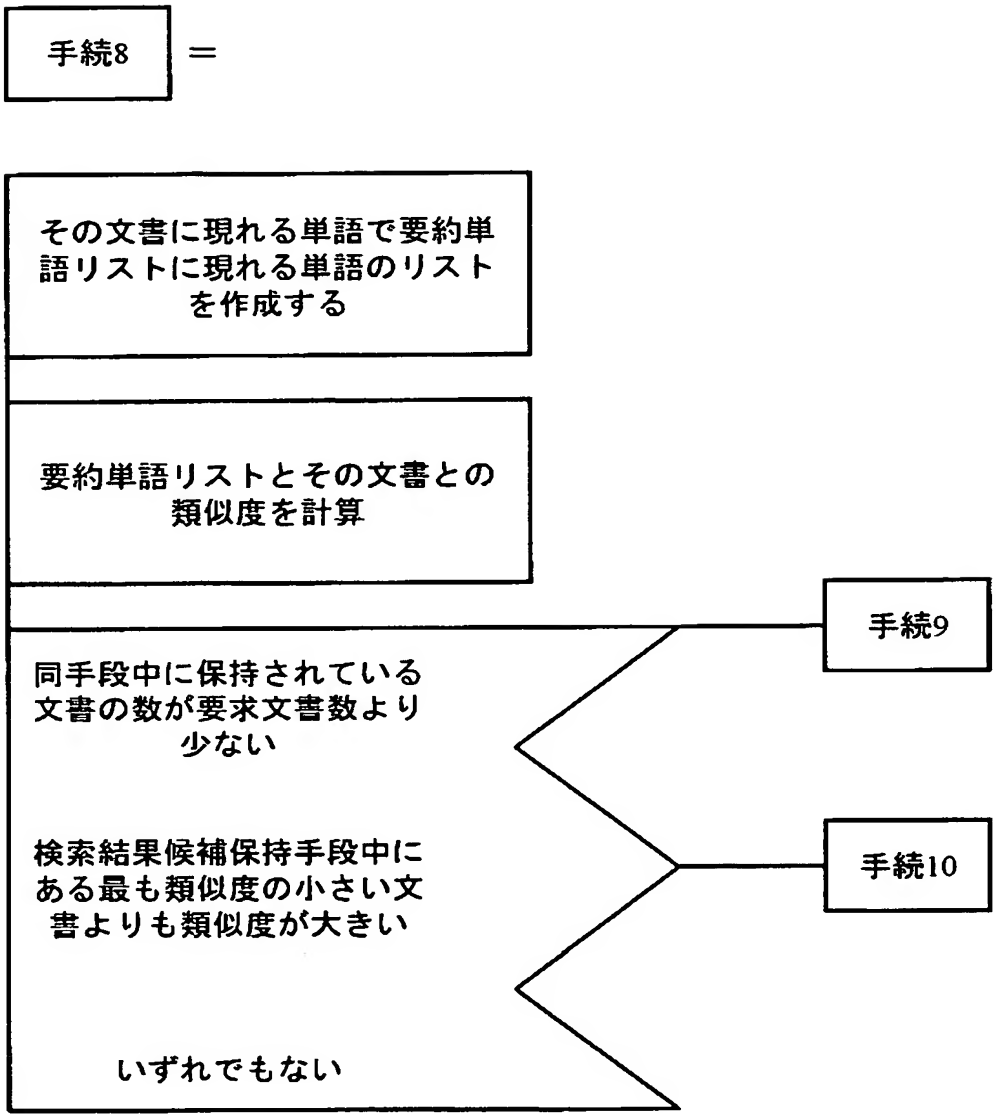
【図１８】

図18


手続7
＝ 要約単語リストから関連する文書を検索（その２）
要約単語リストから、各単語に対応して、その単語を含む文書のリストを作成する
文書のリストに現れる全ての文書について繰り返し
手続8
検索結果候補保持手段に保持されている文書を出力

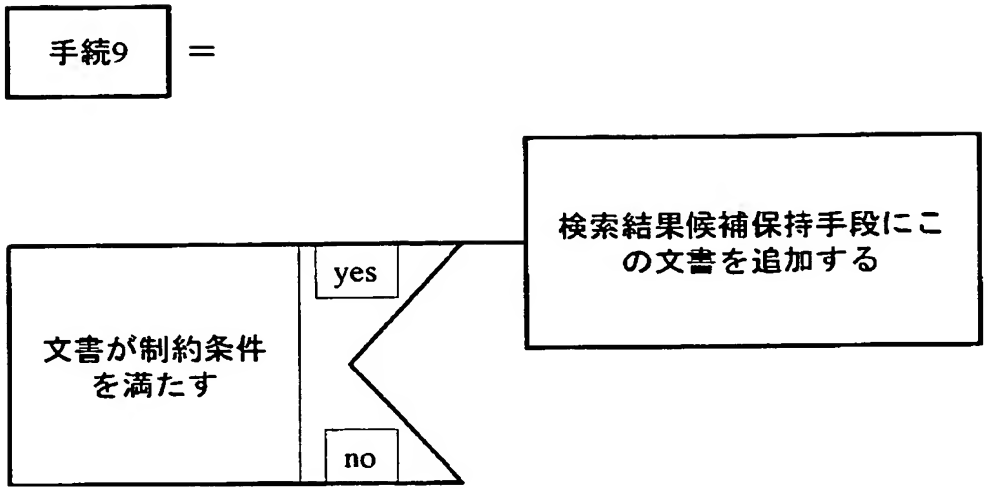
【図１９】

図19


手続8
＝
その文書に現れる単語で要約単語リストに現れる単語のリストを作成する
要約単語リストとその文書との類似度を計算
手続9
同手段中に保持されている文書の数が要求文書数より少ない
手続10
検索結果候補保持手段中にある最も類似度の小さい文書よりも類似度が大きい
いずれでもない

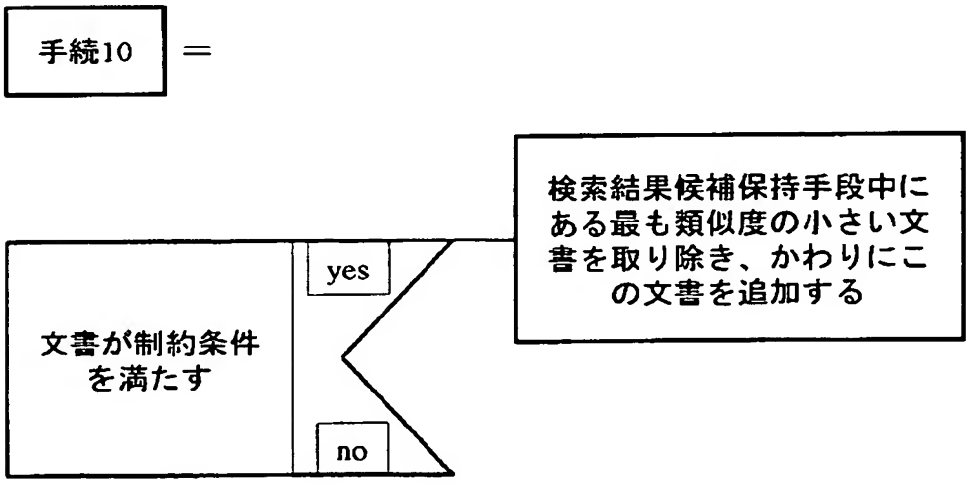
【図２０】

図20



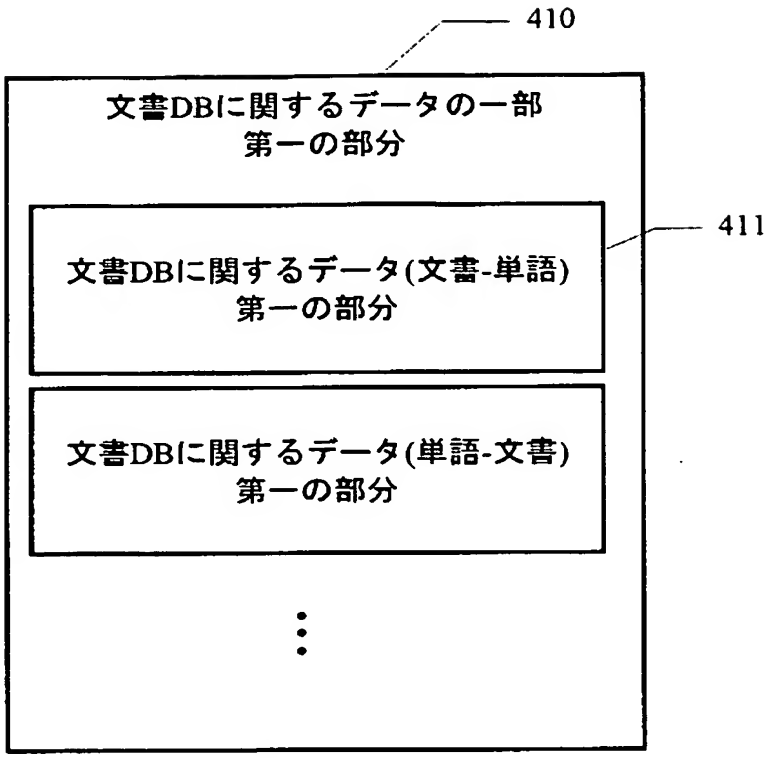
【図２１】

図21



【図２２】

図22


410
文書DBに関するデータの一部
第一の部分
411
文書DBに関するデータ(文書-単語)
第一の部分
文書DBに関するデータ(単語-文書)
第一の部分

【図２３】

図23

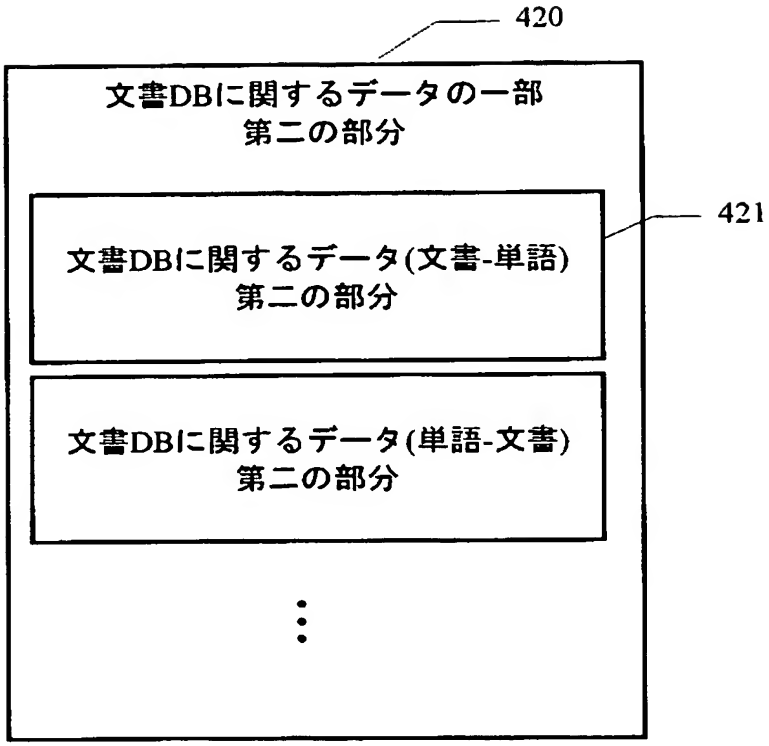
411

文書ID	含まれる単語IDと頻度の対
1	{1, 10}, {2, 3}, ...
2	{2, 5}, ...
3	...
4	...
...	...

この表に現れる単語ID: 1, 2, ...

【図２４】

図24


420
文書DBに関するデータの一部
第二の部分
421
文書DBに関するデータ(文書-単語)
第二の部分
文書DBに関するデータ(単語-文書)
第二の部分

【図２５】

図25

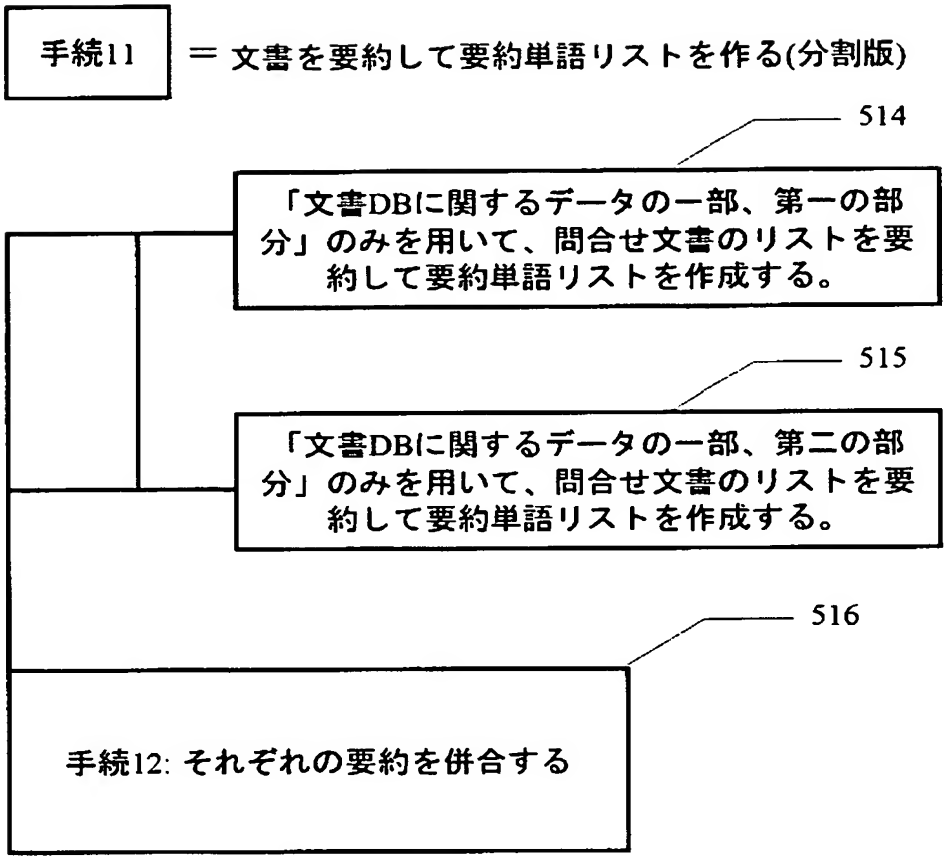
421

文書ID	含まれる単語IDと頻度の対
1	{4, 8}, ...
2	...
3	{4,2},...
4	...
...	...

この表に現れる単語ID: 3, 4, ...

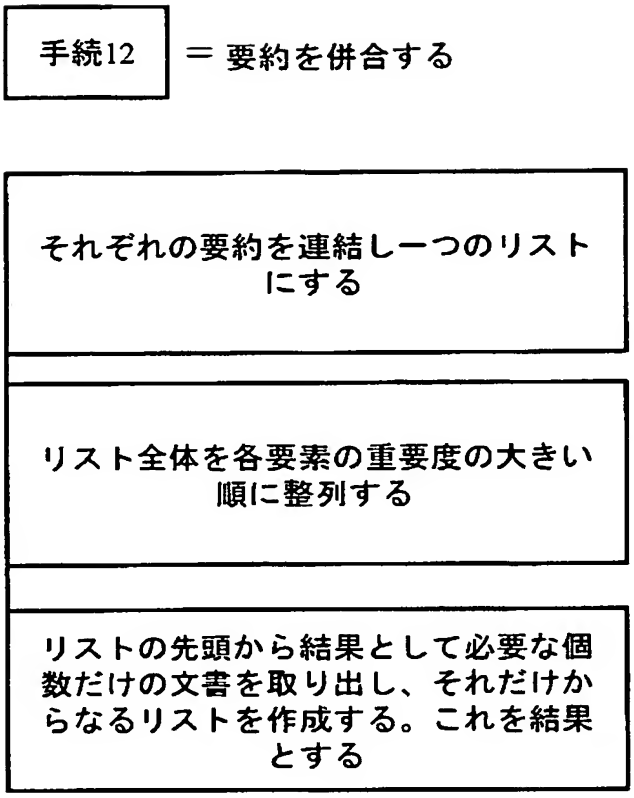
【図２６】

図26


手続11
＝文書を要約して要約単語リストを作る(分割版)
514
「文書DBに関するデータの一部、第一の部分」のみを用いて、問合せ文書のリストを要約して要約単語リストを作成する。
515
「文書DBに関するデータの一部、第二の部分」のみを用いて、問合せ文書のリストを要約して要約単語リストを作成する。
516
手続12: それぞれの要約を併合する

【図２７】

図27



【図２８】

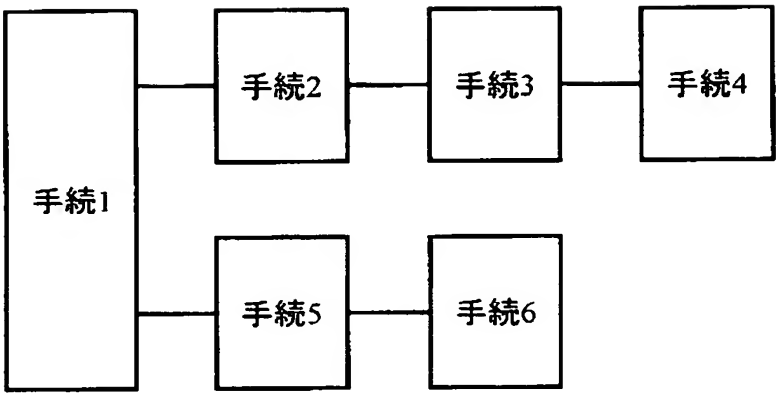
図28

$$e = p\left(1-\sum_{j=0}^{k-1} {}_mC_j\,(1/p)^j\,(1-1/p)^{m-j}\right)$$

分割数をp、要求個数m、各部分が返す個数kとした場合の、
完全な連想・要約結果が得られなくなる確率の上限ｅの計算式

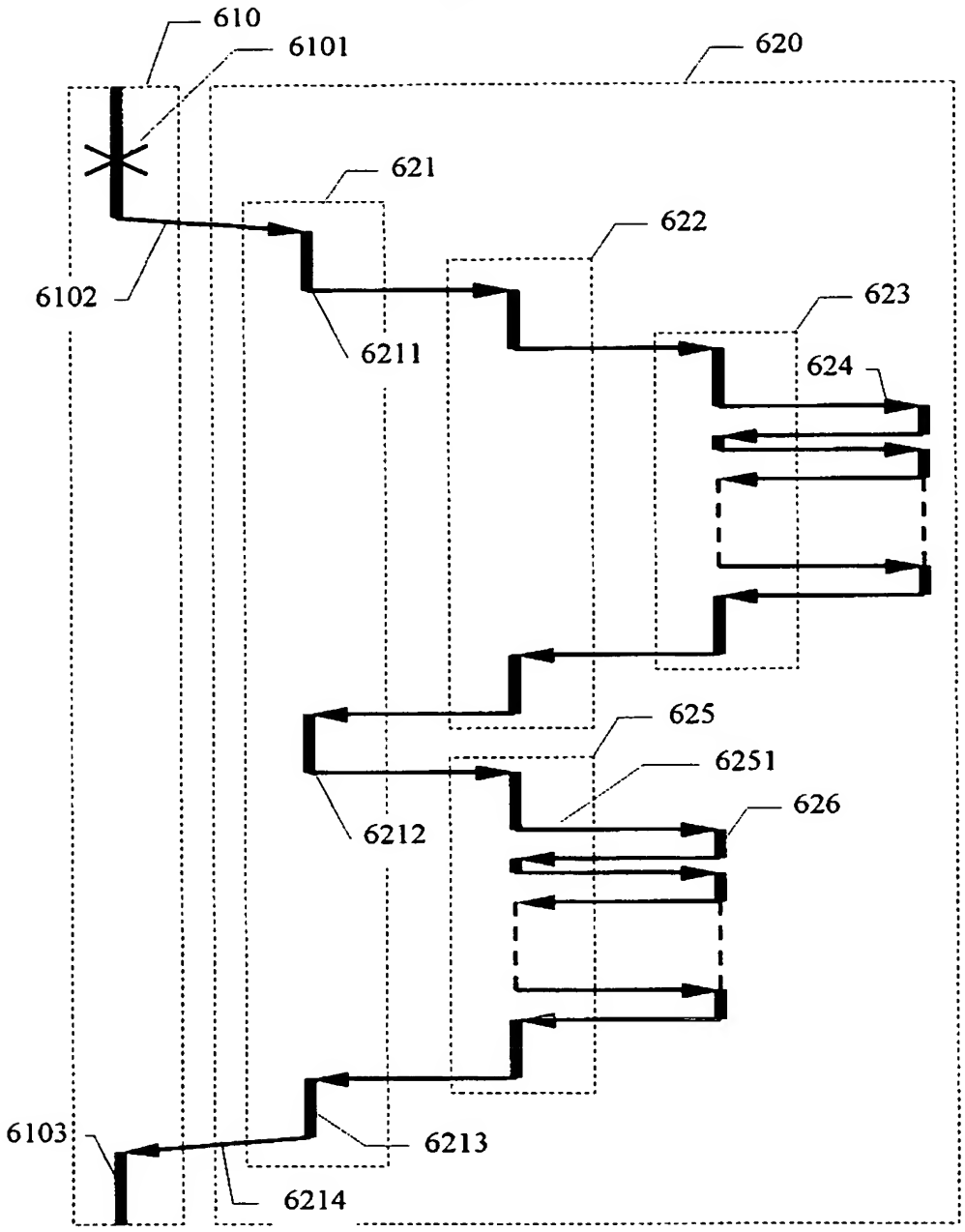
【図２９】

図29



【図３０】

図30


610
6101
620
621
622
6102
623
6211
624
625
6251
626
6212
6103
6213
6214

【書類名】　　要約書

【要約】

【課題】　検索結果の表示順序は、もしくはデータベースに依存した特定の順序で表示されるか、明示的に指定する場合でも何か特定のキーの値で整列させるしか無かったため、検索効率が悪かった。

【解決手段】　検索要求の要約としての要約単語リストを作成し、要約単語リストと検索対象の文書との類似度を計算する類似度計算手段と、検索対象の文書の制約条件を検査する制約条件検査手段とをこの順、またはその逆の順に適用することで、制約条件を満たし、かつ検索要求に類似した文書を検索する。

【効果】　連想検索実行時に制約条件を指定することが可能となり、検索精度や検索作業の効率の向上が可能となる。

【選択図】　図１

認定・付加情報

特許出願の番号	特願２００３－１０４７７１
受付番号	５０３００５８４３２０
書類名	特許願
担当官	第七担当上席　　００９６
作成日	平成１５年　４月１０日

＜認定情報・付加情報＞

【提出日】　平成15年 4月 9日

次頁無

特願２００３－１０４７７１

出　願　人　履　歴　情　報

識別番号　［０００００５１０８］

１．変更年月日　１９９０年　８月３１日
［変更理由］　新規登録
住　所　東京都千代田区神田駿河台４丁目６番地
氏　名　株式会社日立製作所